Sound Source Separation

Evangelista, G; Marchand, S; Plumbley, MD; Vincent, E

Accepted preprint of:
G. Evangelista, S. Marchand, M. D. Plumbley and E. Vincent. Sound source separation. In U. Zölzer (ed.), DAFX: Digital Audio Effects, 2nd edition, Chapter 14, pp. 551-558. John Wiley & Sons, March 2011.
ISBN 978-0-470-66599-2. doi 10.1002/9781119991298.ch14.
Link to article: http://onlinelibrary.wiley.com/doi/10.1002/9781119991298.ch14/summary

Chapter 1

Sound Source Separation

G. Evangelista, S. Marchand, M. D. Plumbley, E. Vincent

1.1 Introduction

When processing a sound recording, sound engineers often face the need to apply specific digital audio effects to certain sounds only. For instance, the remastering of a music recording may require to correct the tuning of a mistuned instrument or relocate that instrument in space without affecting the sound of other instruments. This operation is straightforward when these sounds are available as separate tracks but becomes quite difficult otherwise. Indeed, the digital audio effects reviewed in this book all apply to the recording as a whole.

Source separation refers to the range of techniques aiming to extract the signals of individual sound *sources* from a given recording. The input recording is called *mixture* signal. The estimated source signals can then be separately processed and added back together for remastering purposes. In this scenario, the number of mixture channels is typically equal to one or two or more rarely up to five, while the number of sources ranges from two to ten or more. The need for source separation also arises in many other application scenarios, such as speech enhancement for hearing aids, high-quality upmixing of mono or stereo content to 3D sound formats and automatic speech and speaker recognition in multi-talker environments.

Source separation is a recent field of research compared to the other audio effects reviewed in this book, so that most techniques are less mature and cannot address the above applications scenarios to date. Yet, some established techniques are gradually finding their way to the industry and will soon be part of professional or general consumer software. This chapter will provide an overview of these established techniques as well as more recent ones.

1.1.1 General principles

Notion of source

The first step to address when considering source separation is to formalize the notions of source and mixture. The notion of source or track is often ambiguous in the absence of additional assumptions. For instance, a bass drum, a snare drum and a hi-hat may be considered as separate sources or as components of a single "drums" source depending on the targeted degree of separation. In the following, we make the former choice and assume that all sources are point sources emitting sound from a different point in space. We also set additional constraints on the sources in case of a single-channel mixture.

Modeling of the mixing process

Independently of the application scenario, the notion of mixture can generally be formalized as the result of a multichannel filtering process. Let I be the number of mixture channels and M the number of sources. The point source assumption implies that each source can be represented as a single-channel signal $s_m(n)$, $m \in \{1, \ldots, M\}$. When the sources are digitally mixed by amplitude panning (see Chapter 5), the i-th mixture channel $x_i(n)$, $i \in \{1, \ldots, I\}$, is given by the *instantaneous* mixing model

$$x_i(n) = \sum_{m=1}^{M} a_{im} s_m(n) \tag{1.1}$$

where a_{im} is a scalar panning coefficient. When the mixture is obtained by simultaneous recording of the sources or when additional artificial reverberation is applied, the mixture channels can be expressed by the more general *convolutive* mixing model [SAM07]

$$x_i(n) = \sum_{m=1}^{M} \sum_{\tau} a_{im}(\tau) s_m(n - \tau) \tag{1.2}$$

where $a_{im}(\tau)$ is a Finite Impulse Response (FIR) filter called *mixing filter* modeling time-varying sound transformation between the m-th source and its contribution to the i-th channel[1]. In a conventional recording, the mixing filters are room impulse responses reflecting the spatial directivity of the sources and the microphones and acoustic propagation from the sources to the microphones, including sound reflections over the walls or other objects in the room. The length of the filters is then on the order of a few hundred millisecond in a small room or one second in a concert room. Additional filtering due to the listener's head arises in binaural recordings, so that the mixing filters are equal to the sum of the Head-Related Transfer Functions (HRTFs) associated with the direct sound and with all reflections.

Time-frequency domain processing

Despite its accurate reproduction of the mixing process, the time-domain signal representation (1.2) is generally considered as inconvenient for sound source separation, since all sources except silent ones contribute to each sample $x_i(n)$ of the mixture signal. Time-frequency representations are preferred, since they decrease the overlap between the sources and simplify the modeling of their characteristics. Indeed, all sources are typically characterized by distinct pitches or spectra in a given time frame and by distinct dynamics over time, so that one or two predominant sources typically contribute to each time-frequency bin. This property known as *sparsity* makes it easier to estimate their contributions and subsequently separate them. This phenomenon is illustrated in Fig.1.1 where the features of a violin source and a piano source are hardly visible in the mixture signal in the time domain, but are easily segregated in the time-frequency domain.

Although perceptually motivated representations have been employed [WB06], the most popular time-frequency representation is the Short Time Fourier Transform (STFT), also

[1] Models (1.1) and (1.2) both assume that the source positions and the digital audio effects applied upon them are fixed throughout the duration of the recording. Moving source scenarios have received less attention so far. However, such scenarios may be addressed to a certain extent by partitioning the recording into time intervals over which the mixing filters are reasonably time-invariant and applying the techniques reviewed in this chapter to each interval.

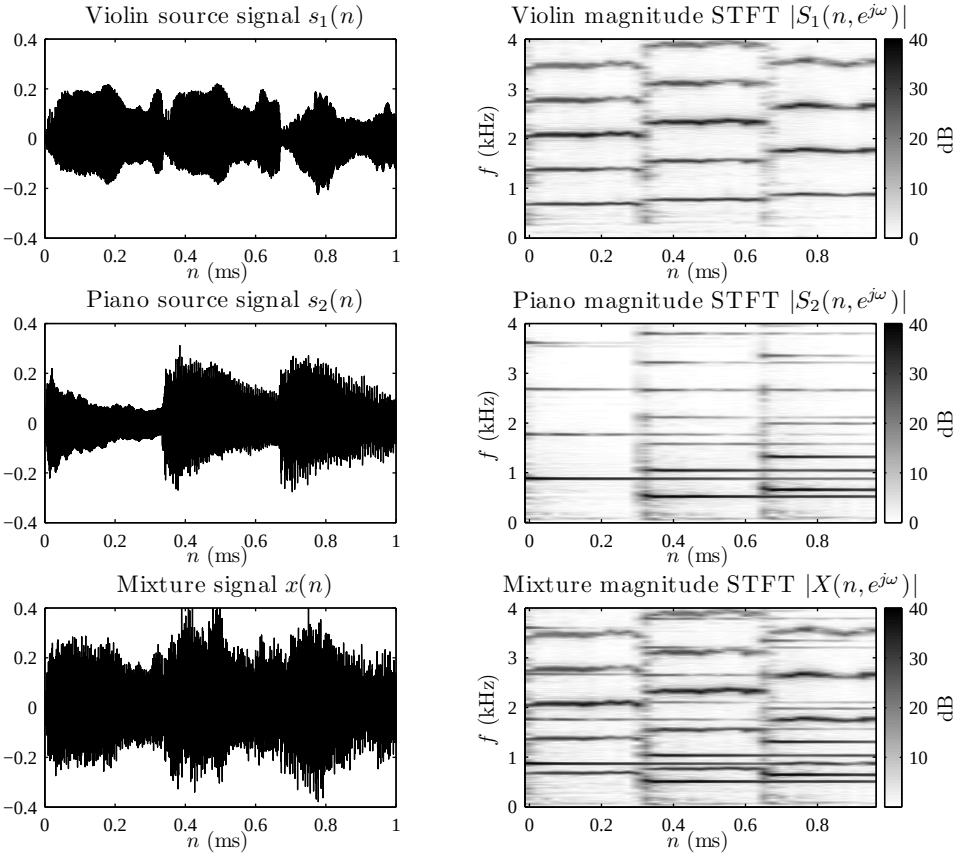


Figure 1.1 Time *vs.* time-frequency representation of a single-channel mixture of a violin source and a piano source.

known as the phase vocoder, that is the subject of Chapter 7. The mixing process can then be modeled directly over the STFT coefficients using the time-frequency filtering algorithm described in that chapter. However, this exact algorithm requires the window length N to be larger than twice the length of the mixing filters, that is typically on the order of several hundred millisecond, while the optimal window length in terms of sparsity is on the order of 50 ms [YR04]. Also it introduces a dependency between the source STFT coefficients at different frequencies due to zero-padding that makes them more complex to estimate. Approximate time-frequency filtering based on circular convolution is hence used instead. Denoting by $X_i(n, e^{j\omega})$ and $S_m(n, e^{j\omega})$ the complex-valued STFT coefficients of the i-th mixture channel and the m-th source in time frame n and at normalized frequency ω and by $A_{im}(e^{j\omega})$ the frequency-domain mixing coefficients corresponding to the mixing filter

$a_{im}(\tau)$, the mixing process is approximately modeled as [SAM07]

$$X_i(n, e^{j\omega}) = \sum_{m=1}^{M} A_{im}(e^{j\omega}) S_m(n, e^{j\omega}). \tag{1.3}$$

Denoting by $\mathbf{X}(n, e^{j\omega})$ the $I \times 1$ vector of mixture STFT coefficients and by $\mathbf{S}(n, e^{j\omega})$ the $M \times 1$ vector of source STFT coefficients, the mixing process can be equivalently expressed as

$$\mathbf{X}(n, e^{j\omega}) = \mathbf{A}(e^{j\omega}) \mathbf{S}(n, e^{j\omega}). \tag{1.4}$$

where $\mathbf{A}(e^{j\omega})$ is the $I \times J$ matrix of mixing coefficients called *mixing matrix*. Source separation amounts to estimating the STFT coefficients $S_m(n, e^{j\omega})$ of all sources and transforming them back to the time domain using overlap-add STFT resynthesis.

Quality assessment

Before presenting actual source separation techniques, let us briefly introduce the terms that will be used to describe the quality of the estimated sources in the rest of this chapter. In practice, perfect separation is rarely achieved, *e.g.* because the assumptions behind source separation algorithms are not exactly met in real-world situations. The level of the target source is then typically increased within each estimated source signal, but distortions remain compared to the ideal target source signals. One or more types of distortion can arise depending on the algorithm [VGF06]: linear or nonlinear distortion of the target source such as *e.g.* missing time-frequency regions, remaining sounds from the other sources, and additional artifacts taking the form of time- and frequency-localized sound grains akin to those observed in denoising applications (see Chapter 7). These three kinds of distortion will be called *target distortion*, *interference* and *musical noise*, respectively. Minimizing interference alone often results in increasing musical noise, so that a suitable trade-off must be seeked depending on the application. For instance, musical noise is particularly annoying and must be avoided at all costs in hearing aid applications.

1.1.2 Beamforming and Frequency-Domain Independent Component Analysis

Separation via unmixing filter estimation

One of the earliest paradigms for source separation consists of estimating the sources by applying a set of appropriate multichannel *unmixing filters* to the mixture. In the time-frequency domain, this amounts to computing the estimated STFT coefficients $\widehat{S}_m(n, e^{j\omega})$ of the sources as [BW01]

$$\widehat{S}_m(n, e^{j\omega}) = \sum_{i=1}^{I} W_{mi}(e^{j\omega}) X_i(n, e^{j\omega}) \tag{1.5}$$

where $W_{mi}(e^{j\omega})$ are complex-valued unmixing coefficients. With similar notations to above, this can be expressed in matrix form as

$$\widehat{\mathbf{S}}(n, e^{j\omega}) = \mathbf{W}(e^{j\omega}) \mathbf{X}(n, e^{j\omega}) \tag{1.6}$$

where $\mathbf{W}(e^{j\omega})$ is the $J \times I$ *unmixing matrix*.

These filters act as spatial filters that selectively enhance or attenuate sounds depending on their spatial position. In order to understand how these filters can be designed to achieve source separation, let us consider at first the simple case of a two-channel mixture of two sources recorded from omnidirectional microphones. The extension of this approach to more than two channels and two sources is discussed later in Section 1.1.3. Since each interfering source generates a large number of echoes at distinct positions, sounds coming from all of these positions should be canceled. Under the assumption that these positions are far from the microphones relatively to the wavelength, sound from a given position will arrive at the second microphone with little attenuation relatively to the first microphone but a delay δ that is approximately equal to

$$\delta = \frac{d\, f_S}{c} \cos\theta \text{ samples} \tag{1.7}$$

where d is the microphone spacing in m, f_S the sampling frequency in Hz, c the speed of sound *i.e.* about $344\,\mathrm{m\,s^{-1}}$ and θ the sound *direction of arrival* (DOA) relative to the microphone axis oriented from the second to the first microphone. Since the frequency response associated with delay δ is equal to $e^{-j\omega\delta}$, the *directivity pattern* of the unmixing filters $W_{mi}(e^{j\omega})$ associated with source p, that is their magnitude response to a sound of normalized frequency ω with DOA θ, is given by [BW01]

$$G_m(\theta,\omega) = \left| W_{m1}(e^{j\omega}) + W_{m2}(e^{j\omega}) e^{-j\omega \frac{d\, f_S}{c} \cos\theta} \right|. \tag{1.8}$$

Note that distance or elevation do not enter into account, provided that the distance is large enough, so that all sounds located on the spatial cone corresponding to a given DOA are enhanced or attenuated to the same extent.

Beamforming

Expression (1.8) allows it to design the filters so as to achieve suitable directivity patterns. Let us assume for a moment that the DOAs θ_1 and θ_2 of both sources are known and that we aim to extract the first source. A first simple design consists of setting

$$W_{11}(e^{j\omega}) = \frac{1}{2} \tag{1.9}$$

$$W_{12}(e^{j\omega}) = \frac{1}{2} e^{j\omega \frac{d\, f_S}{c} \cos\theta_1}. \tag{1.10}$$

This design called *delay-and-sum beamformer* [BW01] adjusts the delay between the two microphone signals before summing them so that they are perfectly in phase for sounds with DOA θ_1 but tend to be out of phase for sounds with other DOAs. The directivity pattern corresponding to this design is depicted in Fig. 1.2. Sounds within a beam around the target DOA are enhanced. However, the width of this beam increases with decreasing frequency and it covers the whole space in the range below 400 Hz with the considered microphone spacing of $d = 30\,\mathrm{cm}$. Sidelobe beams also appear with a regular spacing at each frequency, so that the interfering source is cancelled at certain frequencies only. This results in an average enhancement of 3 dB.

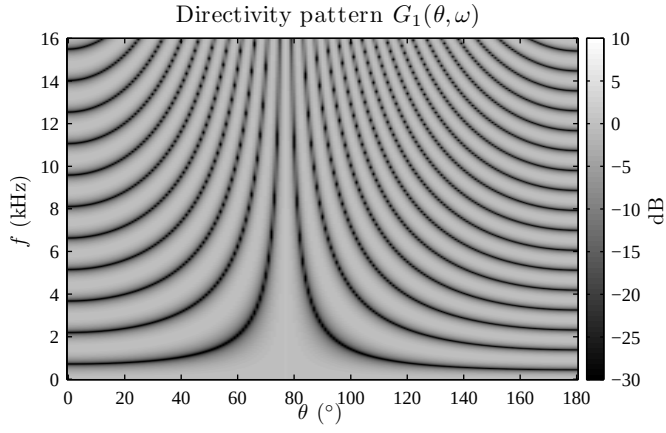


Figure 1.2 Directivity pattern of the two-channel delay-and-sum beamformer pointing to a target source at $\theta_1 = 77°$ with a microphone spacing of $d = 30\,\text{cm}$.

A more efficient design called *null beamformer* [BW01] consists of setting

$$W_{11}(e^{j\omega}) = \frac{1}{1 - e^{j\omega \frac{d\, f_S}{c}(\cos\theta_2 - \cos\theta_1)}} \tag{1.11}$$

$$W_{12}(e^{j\omega}) = -\frac{e^{j\omega \frac{d\, f_S}{c} \cos\theta_2}}{1 - e^{j\omega \frac{d\, f_S}{c}(\cos\theta_2 - \cos\theta_1)}}. \tag{1.12}$$

The numerator of this expression adjusts the delay between the two microphone signals so that sounds with DOA θ_2 are in antiphase, while the denominator adjusts the gain so as to achieve a flat response to sounds with DOA θ_1. The resulting directivity pattern shown in Fig. 1.3 confirms that direct sound from the interfering source is now perfectly notched out, while direct sound from the target source is not affected. Note that the notch at θ_2 is extremely narrow so that precise knowledge of θ_2 is crucial. Also, sidelobes still appear so that echoes of the interfering source are not canceled. Worse, sounds from almost all DOAs are strongly enhanced at frequencies that are multiples of 1.8 kHz with the considered microphone spacing and source DOAs. Indeed, both source DOAs result in similar phase differences between microphones at these frequencies, *i.e.* $\omega\, d\, f_S / c \cos\theta_1 = \omega\, d\, f_S / c \cos\theta_2 \pmod{2\pi}$, so that the numerator tends to cancel the target source together with the interfering source and a strong gain must be applied via the denominator to compensate for this. Precise knowledge of θ_1 is therefore also crucial, otherwise the target source might become strongly enhanced or attenuated at nearby frequencies.

The delay-and-sum beamformer and the null beamformer are both fixed designs, which do not depend on the data at hand except from the source DOAs. More robust adaptive designs have been proposed to attenuate echoes of the interfering source together with its direct sound. For example, the *Linearly Constrained Minimum Variance (LCMV) beamformer* [BW01] minimizes the power of the source signals estimated via (1.5), which is equal to the power of direct sound from the target plus that of echoes and interference,

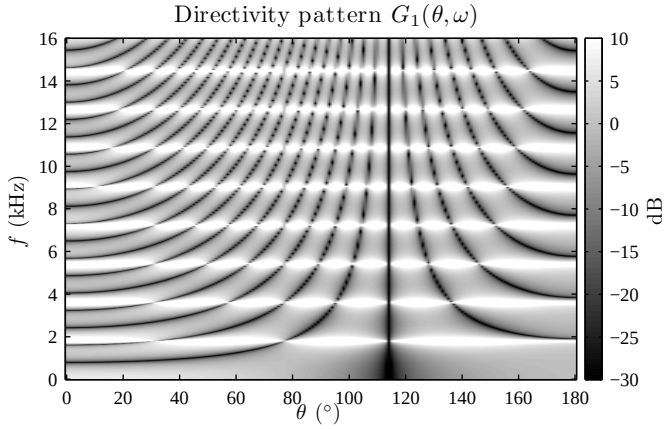


Figure 1.3 Directivity pattern of the two-channel null beamformer for a target source at $\theta_1 = 77^\circ$ and an interfering source at $\theta_2 = 114^\circ$ with a microphone spacing of $d = 30\,\text{cm}$.

while guaranteeing a flat response over the target DOA. This beamformer can be interpreted in a statistical framework as achieving Maximum Likelihood (ML) estimation of the sources under the assumption that the sum of all interfering sounds has a stationary Gaussian distribution. Its implementation via the so-called Generalized Sidelobe Canceller (GSC) [BW01] algorithm does not necessitate knowledge of the interfering source DOA anymore, but still requires precise knowledge of the target DOA. In realistic scenarios, this information is not available and must be estimated from the mixture signal at hand. State-of-the-art source localization algorithms *e.g.* [NSO09] are able to address this issue in anechoic environments, but their accuracy drops in moderately to highly reverberant environments so that the separation performance achieved by beamforming drops as well.

Frequency-Domain Independent Component Analysis

In order to understand how to circumvent this seemingly bottleneck issue, let us come back to the matrix expression of the mixing and unmixing processes in (1.4) and (1.6). By combining these two equations, we get $\widehat{\mathbf{S}}(n, e^{j\omega}) = \mathbf{W}(e^{j\omega})\mathbf{A}(e^{j\omega})\mathbf{S}(n, e^{j\omega})$. Therefore, if the mixing filters were known, choosing the unmixing coefficients as

$$\mathbf{W}(e^{j\omega}) = \mathbf{A}(e^{j\omega})^{-1} \tag{1.13}$$

would result in perfect separation *i.e.* $\widehat{\mathbf{S}}(n, e^{j\omega}) = \mathbf{S}(n, e^{j\omega})$ in the limit where time-frequency domain approximation of the mixing process is valid and $\mathbf{A}(e^{j\omega})$ is invertible. In practice, $\mathbf{A}(e^{j\omega})$ can be singular or ill-conditioned at the frequencies for which the sources result in similar phase and intensity differences between microphones. The directivity pattern corresponding to these optimal unmixing coefficients is illustrated in Fig. 1.4 in the case of a concert room recording. Deviations compared to Fig. 1.3 are clearly visible and due to summation of direct sound and echoes at the microphones, resulting in apparent DOAs different from the true DOAs.

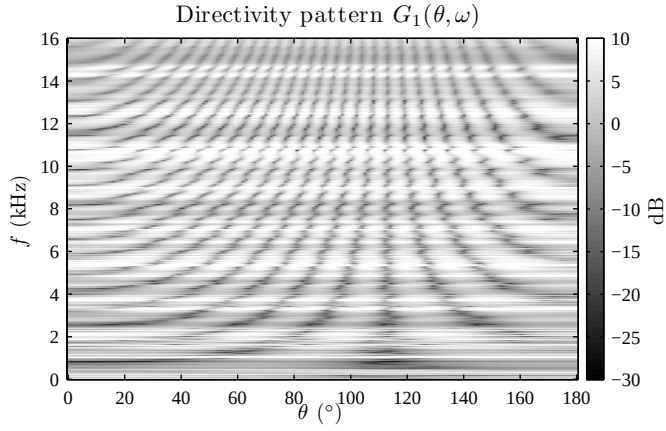


Figure 1.4 Directivity pattern of the optimal unmixing coefficients for a target source at $\theta_1 = 77^\circ$ recorded in a concert room in the presence of an interfering source at $\theta_2 = 114^\circ$ with a microphone spacing of $d = 30\,\text{cm}$.

In practice, the mixing filters are unknown, thus the optimal unmixing coefficients must be adaptively estimated from the mixture signal. This can be achieved in a statistical framework by ML estimation of the unmixing coefficients under the assumption that the STFT coefficients of all sources are independent and follow a certain distribution. It can be shown that the ML objective is equivalent to maximizing the statistical independence of the STFT coefficients of the sources, hence this approach is known as *Frequency-Domain Independent Component Analysis (FDICA)*. A range of prior distributions have been proposed in the literature [SAM07, VJA+10] which typically reflect the aforementioned sparsity property of the sources, *i.e.* the fact that the source STFT coefficients are significant in a few time frames only within each frequency bin. Note that this statistical framework is very different from that underlying LCMV beamforming, since the sources are now modeled as separate sparse variables instead of a joint Gaussian "noise".

A popular family of distributions is the *circular generalized Gaussian* family [GZ10]

$$P(|S_m(n, e^{j\omega})|) = \frac{p}{\beta\,\Gamma(1/p)} \exp\left(-\left|\frac{S_m(n, e^{j\omega})}{\beta}\right|^p\right) \tag{1.14}$$

where $\Gamma(.)$ denotes the function known in mathematics as the gamma function. The scale and shape parameters β and p govern respectively the average magnitude and the sparsity of the source STFT coefficients. The smaller p, the more coefficients concentrate around zero. Distributions with shape parameter $p < 2$ are generally considered as sparse and those with $p > 2$ as non-sparse with respect to the Gaussian distribution over the magnitude STFT coefficients associated with $p = 2$. In the absence of prior information about the spectral shape of the sources, the scale parameters β is typically fixed so that the coefficients have unit power. Fig. 1.5 shows that this distribution with $p = 0.4$ provides a very good fit of the distribution of a speech source after power normalization in each frequency bin. The shape parameter value $p = 1$, which results in the slightly less sparse *Laplacian distribution*, is nevertheless a popular choice [SAM07].

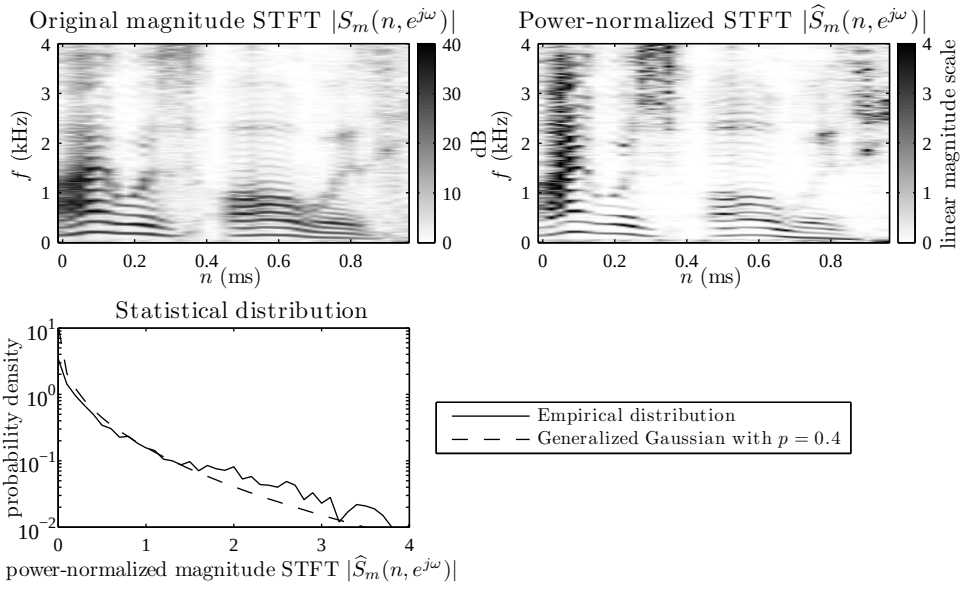


Figure 1.5 Distribution of the power-normalized magnitude STFT coefficients of a speech source compared to the generalized Gaussian distribution with shape parameter $p = 0.4$.

The likelihood of the observed mixture signal is obtained by multiplying the probability density (1.14) over all sources, all time frames and all frequency bins. ML estimation of the unmixing coefficients is then equivalent to solving the following optimization problem:

$$\min_{W_{mi}(e^{j\omega})} \sum_{m=1}^{M} \sum_{n,\omega} |\widehat{S}_m(n, e^{j\omega})|^p \tag{1.15}$$

where $\widehat{S}_m(n, e^{j\omega})$ implicitly depends on $W_{mi}(e^{j\omega})$ via (1.5). This problem may be addressed using a range of algorithms, that rely on principles of optimization theory beyond the scope of this chapter. Readers may refer to [SAM07] for details.

Two additional problems remain. Firstly, since the scale parameter β is constant over all frequency bins, the resulting sources have a flat average spectrum. This *scaling indeterminacy* of FDICA may be circumvented by exploiting more advanced models of the source spectra or by multiplying the estimated source STFT coefficients by the mixing coefficients derived from the unmixing coefficients via (1.13) so as to estimate the contribution of each source to one or more mixture channels instead of the original source. Secondly, since the model (1.14) is identical for all sources, the sources can be estimated at best up to a random order. This *permutation indeterminacy* of FDICA can be addressed to a certain extent by exploiting more advanced models of the source spectra or by estimating approximate source DOAs and permuting the unmixing coefficients so that the resulting directivity patterns match the null beamformer pattern in Fig.1.3 as closely as possible. Again, see [SAM07]

for details.

1.1.3 Statistically Motivated Approaches for Under-Determined Mixtures

We have seen that null beamforming or preferably FDICA can be employed to separate a two-channel mixture of two sources. At this stage, most readers will undoubtedly wonder how these methods generalize to more channels and more sources. As it turns out, these algorithms can be extended to any number of sources and channels. However, they are efficient only when the number of sources is equal or smaller than the number of channels. Indeed, the number of independent spatial notches corresponding to any set of unmixing filters is at most equal to the number of channels. This can be seen in Fig.1.3 where the unmixing coefficients may be chosen so as to form one notch per frequency in the direction of the interfering source in a two-channel mixture but other notches due to sidelobes cannot be independently adjusted. Mixtures satisfying this constraint are called *determined mixtures*.

In the rest of this chapter, we shall focus on *under-determined mixtures*, which involve more sources than mixture channels and occur more often in practice. The separation of such mixtures requires a novel paradigm: instead of estimating unmixing filter coefficients, one now wants to estimate the mixing coefficients and the source STFT coefficients directly. Again, this can be addressed in a statistical framework by specifying suitable statistical models for the mixing coefficients and the source STFT coefficients. In practice, for historical reasons, most methods rely on a statistical model of the source STFT coefficients but adopt a simpler deterministic model for the mixing coefficients based on *e.g.* perceptual considerations.

Two categories of models have been studied so far, depending on the number of channels. In a multichannel mixture, spatial information still helps separating the sources so that weak source models are sufficient. Sparse distributions of the source STFT coefficients have been used in context together with learned mapping functions between the source DOAs and the mixing coefficients. An example algorithm relying on such distributions will be presented in Section 1.2. In a single-channel mixture, spatial information is no more available so that more accurate models of the source STFT coefficients are needed. Example algorithms relying on such models will the described in Section 1.3.

1.1.4 Perceptually Motivated Approaches

The ability to locate and separate sound sources is well developed among humans and animals who use this feature to orient themselves in the dark or to detect the potential sources of danger. The field of *Computational Auditory Scene Analysis (CASA)* studies artificial systems able to mimic this localization-separation process. These approaches clearly represent a shift in the paradigm where rather than in modeling the sound production process focus is in modeling the spatial sound perception processes.

A simple task in source localization is to detect, to a good degree of approximation, the DOA of the source waves from the signals available at the two ears. To a certain extent, humans are able to perform this task even if one of the two ears is obstructed or not functioning, i.e., from monaural hearing. Several models have been conceived to explain binaural perception, which are rooted in the work by Jeffress [JSY98] on the analysis of *Interaural Time Difference* (ITD) by means of a neuronal coincidence structure, where nerve impulses derived from each of the two ears stimuli travel in opposite directions over delay lines. This model transforms time information into space information since a nerve

cell is excited only at the position on the delay lines where the counter-traveling impulses meet.

Back at the beginning of last century, the ITD together with the *Interaural Level Difference* (ILD) were considered as the principal spatial hearing cues by Lord Rayleigh, who developed the so called *Duplex Theory of Localization*. According to this theory, validated by more recent experimentation, ITDs are more reliable at lower frequencies (roughly below 800 Hz), while ILDs perform better at higher frequencies (roughly above 1.6 kHz). This is due to the fact that the wavelength associated with low audio frequencies is larger than the distance between the ears (typically 12-20 cm). In this case, the perturbation at the two ears is mainly characterized by phase differences with almost equal levels. On the other hand, at higher frequencies, the head is shadowing the perturbation reaching one of the ears, thus introducing relevant ILDs for sound sources that are not placed directly in the frontal direction of the listener. In real measurements, ITDs and ILDs are the result of multiple reflections, diffractions and resonances generated in the head, torso and external ears of the listener. Consequently, the interpretation and use of ITD and ILD as cues for DOA detection is less simple and error prone [Bla01, Gai93, Lin86, ST97].

In perceptually motivated source separation methods, the binaural signals are first input to a cochlear filter bank. The sources are separated by processing linearly or non-linearly the signal in perceptual frequency bands. Typical non-linear processing includes half-wave rectification or squaring followed by low-pass filtering [FM04]. ITD and ILD are also estimated within perceptual bands and used as cues for the separation.

While the accurate modeling of binaural hearing is essential to the understanding of auditory scene analysis as performed by humans or animals, several approaches to sound source localization and separation based on spatial cues were derived for the sole purpose of processing audio signals. In this case, the proposed algorithms are often only crudely inspired by human physiological and perceptual aspects. The ultimate goal is to optimize performance, without necessarily adhering to biological evidence. For example, the frequency resolution at which ITD and ILD cues can be estimated can go well beyond that of the critical frequency bands of the cochlear filter bank.

1.2 Binaural Source Separation

In binaural separation, the various sources are segregated according to spatial cues extracted from the signals available at both ears. This is the case of the signals transduced by the microphones of a two-ear hearing aid system or of the signals available at the microphones in the artificial ears of a dummy head. The type of apparatus is irrelevant, as long as human-like head shadowing is present in-between the microphones. A common strategy is to first detect and discern sources based on the DOA of waves, which is the object of Section 1.2.1, and then to build suitable time-frequency masks in a two-channel STFT representation of the binaural signal in order to demix them. Each mask, which can be binary or continuous valued with arbitrary values in $[0, 1]$, coarsely represents the time-frequency profile of a given source. Due to energy leakage in finite sliding window time-frequency analysis, the masks are bound to cover broader bands than the ideal time-frequency tracks. The estimation of proper masks is the object of Section 1.2.2. The masks are multiplicatively applied to both STFT channels and the transforms inverted. The ideal outcome is a set of binaural signals, each representing the sound of a single source spatially placed at the corresponding DOA.

1.2.1 Binaural Localization

An important aspect of binaural localization is the estimation of the DOA in terms of azimuth and elevation. Together with the range information (distance from the source), it provides full knowledge of the coordinates of the source. However, we shall confine ourselves in estimates of the azimuth only. Estimates of the range are difficult in closed spaces and if the distance of the source to the listener is not large. Estimates of the elevation are affected by larger error, even in the human hearing system. The so called cones of confusion [Bre90], show large uncertainty on the elevation of the source, while azimuth uncertainty as sharp as $\pm 5^{\circ}$ is common in humans. It must be pointed out that the azimuth resolution depends on the angle, with lateral directions providing less sharp results.

In order to estimate the azimuth of the source one can explore spatial cues such as ILD and ITD. For any given source and at any given time, the ILD is the difference in level as received at the ears. Given the STFT of the left and right ear signals, $X_L\left(n, e^{j\omega}\right)$ and $X_R\left(n, e^{j\omega}\right)$, respectively, one can estimate the ILD at any angular frequency ω and time interval indexed by n, where the signal energy is above a certain threshold, as follows:

$$\Delta L_n(\omega) = 20 \log_{10} \left| \frac{X_R\left(n, e^{j\omega}\right)}{X_L\left(n, e^{j\omega}\right)} \right|. \tag{1.16}$$

Basically, the ILD estimate is the difference in dB of the amplitudes in the two channels as a function of frequency and averaged over the STFT analysis interval $[nN, ..., nN + M - 1]$, where the integers N and M respectively denote hop-size and window length of a discrete time STFT.

For any given source and at any given time, the ITD is defined as the time shift of the two signals received at the ears. This time shift could be measured, e.g., by finding a local maximum of the two signals cross-correlation. However, in the time-frequency paradigm we adopted for the whole localization-separation process, for each angular frequency ω one can estimate the ITD from the STFT of the two ear signals, as follows:

$$\Delta T_{n,p}(\omega) = \frac{1}{\omega} \left(\angle \frac{X_R\left(n, e^{j\omega}\right)}{X_L\left(n, e^{j\omega}\right)} + 2\pi p \right) \tag{1.17}$$

The ITD is estimated from the phase difference of the two ear signals. In fact, through division by ω, it corresponds to a phase delay measurement. Since the phase is defined modulo 2π, there is an ambiguity in the estimate that is a multiple of this quantity, which justifies the term $2\pi p$ in (1.17), where p is any integer.

By the nature of the problem, ITD estimates are sharper at lower frequencies, where the effect of phase ambiguity is minimized, while ILD estimates are sharper at high frequencies, i.e., at wavelengths much shorter than the distance between the ears.

Localization using HRTF data or head models

Azimuth estimation from ITD and ILD can be performed by means of table lookup, using a set of measured HRTFs for a given individual. Given the measurements $HRTF_R^s(\theta, e^{j\omega})$ and $HRTF_L^s(\theta, e^{j\omega})$ of the subjective HRTF for the right and left ear as a function of the azimuth θ and angular frequency ω, one can form the following tables

$$\Delta L_s(\theta, e^{j\omega}) = 20 \log_{10} \left| \frac{HRTF_R^s(\theta, e^{j\omega})}{HRTF_L^s(\theta, e^{j\omega})} \right| \tag{1.18}$$

and

$$\Delta T_{s,p}(\theta, e^{j\omega}) = \frac{1}{\omega}\left(\angle \frac{HRTF^s_R(\theta, e^{j\omega})}{HRTF^s_L(\theta, e^{j\omega})} + 2\pi p\right) \tag{1.19}$$

for the azimuth lookup, respectively, from ILD and ITD estimates. As shown in Fig. 1.6 left column, the measured HRTF tables are quite noisy, depending on the details of multiple head-torso wave reflection. It is practice to smooth these tables along the azimuth axis, as shown in Fig. 1.6 right column; we will continue to use the same symbols as in (1.18) and (1.19) to denote the smoothed tables.

For any frequency, given an STFT based estimate $\Delta L_n(\omega)$ for the level difference as in (1.16), the azimuth θ can be estimated by finding in (1.18) which value of θ provides the measured ILD. We denote this estimate by $\theta_{L,n}(\omega)$.

Similarly, given an STFT based estimate $\Delta T_{n,p}(\omega)$ for the time difference as in (1.17), the azimuth can be estimated by finding in (1.19) which value of θ provides the measured ITD. The time difference estimate depends on an arbitrary integer multiple p of 2π. Therefore, for any fixed frame index n and angular frequency ω, each ITD estimate is compatible with a countable infinity of azimuth estimates $\theta_{T,n,p}(\omega)$. However, assuming that the phase difference of the HRTFs does not show large discontinuities across azimuth and that the phase difference at zero azimuth is as small as possible, i.e. zero, it is possible to resolve the phase ambiguity by unwrapping the phase difference of the right and left HRTFs along the azimuth. Even when this ambiguity is resolved, there can still be several values of θ providing the same ILD or ITD values in HTRF lookup. While smoothing tends to partially reduce ambiguity, majority rules or statistical averages over frequency or other assumptions can be used to increase the reliability of the azimuth estimate.

Usually, the Duplex Theory is applied here to choose among the estimates as a function of frequency. At low frequencies, the estimate $\Delta T_{s,0}(\theta, e^{j\omega})$ is selected for the azimuth lookup estimates from unwrapped phase difference, while at high frequencies $\Delta L_s(\theta, e^{j\omega})$ is selected. In Section 1.2.1 a procedure for azimuth estimate jointly employing ITD and ILD is described.

Given a head radius of r meters, the cut-off frequency f_c to switch from ITD to ILD based azimuth estimates can be estimated as the inverse of the time for the sound waves to travel from ear to ear along the head semicircle, i.e., $f_c = \frac{c}{\pi r}$, where $c \approx 344$ ms^{-1} is the speed of sound in air. Approximately a cut-off frequency of 1.6 kHz is selected to switch from ITD to ILD based estimates when the head radius is not known.

The discussed azimuth estimation lookup procedure requires knowledge of the individual HRTFs. When these are not known, it is possible to resort to a slightly less accurate estimation procedure which makes use of average ILD and ITD, respectively obtained from (1.18) and (1.19) by averaging over several individuals in a database of HRTFs. It can be shown [RVE10] that the performance of the averaged model is comparable with that of the method based on individual transfer functions.

At the cost of slightly reduced performance, it is possible to eliminate the need for HRTF based lookup tables in azimuth estimates. In fact, from simple geometric considerations [WS54], as shown in Fig. 1.7 for an idealized face represented by a circular section of radius r, the two ears path length difference of a wave reaching from DOA θ is not only due to an "in air" term $r \sin\theta$ but also to the arc length $r\theta$ of a curved path along the face, yielding

$$\Delta T(\theta) = r\frac{\sin\theta + \theta}{c} \tag{1.20}$$

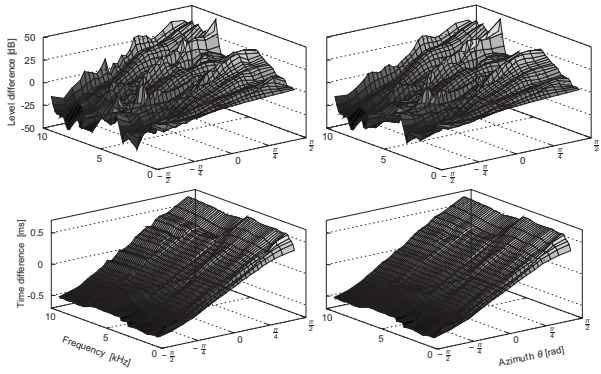


Figure 1.6 Spatial cues from HRTF measurements; Left: Unsmoothed estimates, Right: Estimates smoothed along azimuth axis; Top: Level Differences; Bottom: Time Differences.

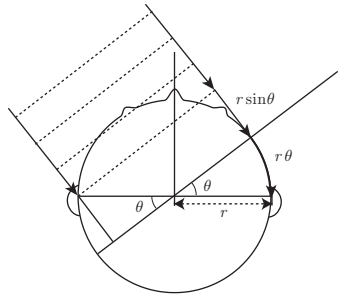


Figure 1.7 Path length difference model for the ITD based on head geometry.

Due to the fact that the head is not perfectly spherical the measured ITD is slightly larger than the values in (1.20), especially at low frequencies. A more accurate model, discussed in [RVE10], scales the ITD model (1.20) by a frequency and, possibly, subject dependent scaling factor $\alpha_s(\omega)$, to obtain

$$\Delta T_s(\theta, \omega) = \alpha_s(\omega) r \frac{\sin\theta + \theta}{c} \tag{1.21}$$

Given a measure of the ITD $\Delta T_s(\theta, \omega)$, in order to produce an estimate for the azimuth θ one needs to invert (1.21), which can be achieved by polynomial approximation [RVE10].

The ILD is a much more complex function of frequency. However, based on the observation of a large number of HRTFs, the following model [RVE10] can be shown to capture its main features when the sources are at large distances (over 1.5 m) from the listener:

$$\Delta L_s(\theta, \omega) = \beta_s(\omega) \sin\theta \tag{1.22}$$

where $\beta_s(\omega)$ is again a frequency and, possibly, subject dependent scaling factor. Inversion of (1.22) in order to produce an estimate for the azimuth does not present any problem.

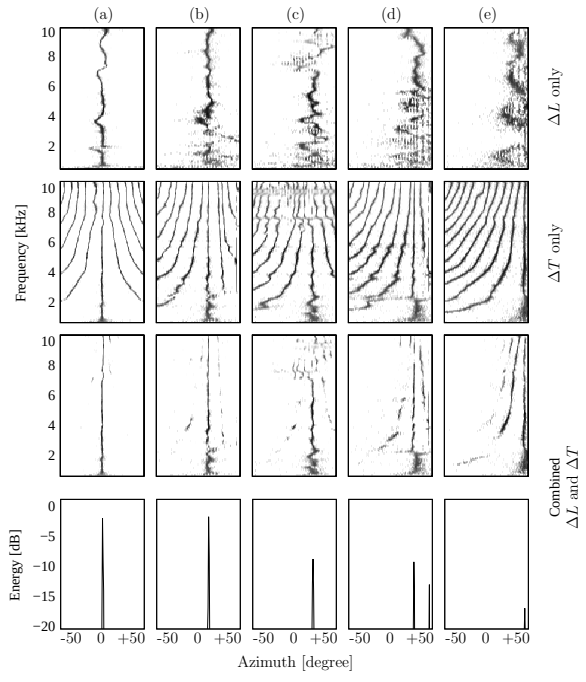


Figure 1.8 2D time histograms of azimuth estimates, in terms of azimuth and frequency, for five different heads and azimuth angles an 0°, 15°, 30°, 45°, and 65°, (a)-(e), respectively. First row: based on ILD only. Second row: based on ITD only. Third row: based on combined evaluation of ILD and ITD. Bottom row: marginal histograms obtained by summing the combined ITD-ILD evaluation over all frequencies.

Localization using ITD, ILD and IC

Since both the ILD and the ITD are related to the azimuth, they can be jointly employed in order to improve the azimuth estimates. The noisy $\Delta L_n(\omega)$ provides a rough estimate of the azimuth for each left/right spectral coefficient pair. This ILD estimate can be used in order to select the most reasonable value for the parameter p in the ITD estimate, which is the one for which the ILD and ITD based estimates provide the closest value for the azimuth. Formally, this is given by

$$\theta_{J,n} = \theta_{T,n,\rho|\rho=\text{argmin}_p\left|\theta_{T,n,p}-\theta_{L,n}\right|} \tag{1.23}$$

Comparative experimental results obtained by ILD only, ITD only and joint method are shown in Fig. 1.8 for a single white noise source located at angles 0°, 15°, 30°, 45° and 65°. Extensive performance evaluation of the joint ILD-ITD azimuth estimates can be found in [RVE10].

In the presence of one source only, the two ears signals x_L and x_R approximately are two delayed and amplitude scaled versions of the same signal. In that case, the *Interaural*

Cross Correlation (ICC), defined as the normalized cross-correlation

$$\rho_{L,R}(n) = \frac{\sum\limits_{m} x_L(m) x_R(n+m)}{\sqrt{\sum\limits_{p} x_L^2(p) \sum\limits_{q} x_R^2(q)}} \tag{1.24}$$

peaks at a time lag corresponding to the relative delay or global ITD of the two signals. The amplitude of the peak provides a measure of similarity. In the presence of two or more sources with different DOA, the cross-correlation peaks are less distinguishable: since there is more than one characteristic delay, the two ear signals are less similar. The *Interaural Coherence (IC)* is defined as the maximum of the normalized cross-correlation of the two ears signals, which is a number between 0 and 1. It is useful to evaluate the ICC in time-frequency, i.e. with $x_L(n)$ and $x_R(n)$ in (1.24) replaced by the STFTs $X_L\left(n, e^{j\omega}\right)$ and $X_R\left(n, e^{j\omega}\right)$, respectively:

$$\rho_{L,R}\left(n, e^{j\omega}\right) = \frac{\left|\sum\limits_{m} X_L\left(m, e^{j\omega}\right) X_R^*\left(n+m, e^{j\omega}\right)\right|}{\sqrt{\sum\limits_{p} \left|X_L\left(p, e^{j\omega}\right)\right|^2 \sum\limits_{q} \left|X_R\left(q, e^{j\omega}\right)\right|^2}} \tag{1.25}$$

where $*$ denotes complex conjugation. In practice, the ICC is smoothed in time [Men10] in order to mitigate the oscillatory behavior due to the STFT finite window length. The corresponding time-frequency IC cue, defined as the amplitude of the maximum of the time-frequency ICC for any fixed frequency, provides a measure of reliability of the ITD and ILD for each frequency channel [FM04], IC being larger if only one source contributes to a given frequency bin. Furthermore, the time-frequency IC can be effectively employed as a cue to improve separation in the presence of time-frequency overlapping sources, where non-binary demixing masks are optimized by maximizing the IC of the separated sources.

1.2.2 Binaural Separation

In Section 1.2.1, methods for the localization of sources in binaural signals are discussed, which are based on DOA discrimination from the STFT of the signals observed at the two ears. The azimuth $\theta(n, \omega)$ estimated for each left/right pair of spectral coefficients pertains to a narrow-band signal component in a short time interval. The ensemble of azimuth estimates can be employed to obtain separate binaural signals where the spatial information is preserved, but where only one source is present in each, which is the object of this section. Starting from the simple method based on binary masks we then explore the construction of countnuous valued masks based on Gaussian mixtures models of the multimodal distribution of azimuth estimates. We then illustrate the use of structural assumption in the multi-channel separation of overlapping harmonics.

Binary Masks

At any given time, each prominent peak of the histogram $h(\theta)$ obtained by cumulating the azimuth estimates $\theta(n, \omega)$ over frequency pertains to different observed DOA. In the assumption that the sound sources are spatially distributed, so that each has a sufficiently distinct DOA as observed by the listener, the number of peak equals the number of sources.

The position of the peak estimates the DOA $\theta_k(n)$ of the corresponding source, where k represents the source index, which can be tracked over time.

Given the azimuth estimates $\theta(n,\omega)$ and the source azimuths $\theta_k(n)$, a weighting factor $M_k(n,\omega)$ can be given for each spectral coefficient. For each source, the separated binaural signal is obtained by multiplying this weighting factor by the STFT of the left and right ear signals followed by STFT inversion. Respectively, the left and right channel STFTs of the k-th reconstructed source signal are given by:

$$\begin{aligned} Y_{k,L}(n,\omega) &= M_k(n,\omega)X_L(n,\omega) \\ Y_{k,R}(n,\omega) &= M_k(n,\omega)X_R(n,\omega) \end{aligned} \tag{1.26}$$

Different approaches can be considered for the definition of the weights. In the assumption that the STFT spectra of the different sources do not overlap significantly, i.e., in the Window-Disjoint Orthogonal (WDO) assumption [YR04]. In this case, only one source contributes significant energy to a given spectral coefficient so that each spectral coefficient can be exclusively assigned to that source. Binary weights can be efficiently employed in this case. A possible strategy is to assign each spectral coefficient to the source whose azimuth estimate is closest. This is formalized by the following choice of weights:

$$M_k(n,\omega) = \begin{cases} 1 & \arg\min_m |\theta(n,\omega) - \theta_m(n)| = k \\ 0 & otherwise \end{cases} \tag{1.27}$$

This weight system can be considered as a spatial window in which the coefficients are assigned to a given source according to the closeness of direction. However, this choice of weights can lead to artifacts due to the fact that azimuth estimates that are very far from the estimated azimuth of any source are unreliable. Therefore, the corresponding STFT coefficients should not be arbitrarily assigned to a source in this case. It may be better to consider these estimates as outliers and assign to a given source only those STFT coefficients for which the corresponding azimuth estimates lie in a window of width W from the estimated source azimuth. This is formalized by the following choice of weights:

$$M_k(n,\omega) = \begin{cases} 1 & |\theta(n,\omega) - \theta_k(n)| < \frac{W}{2} \\ 0 & otherwise \end{cases} \tag{1.28}$$

The result of the separation of three trombone sources, each playing a different note, located at -30°, 15° and 45° are shown in Fig. 1.9, together with the spectrogram displayed in Fig. 1.10. Artifacts are visible and audible especially in the attack area and during silence.

Smoothing in the time direction can effectively reduce artifacts due to mask switching. However, in performed music, more severely than in speech, the overlap of the sources in the time-frequency is not a rare event. In fact, musicians are supposed to play with the same tempo notes that are harmonically related, which means that several partials of the various instruments are shared within the same time interval. This results in artifacts of the separation when binary weighting of either of the forms (1.27) and (1.28) is enforced. One way to reduce these artifacts is to enforce non-binary weighting in which the energy of time-frequency bins is distributed among two or more overlapping sources, with different weights. Even in the ideal case where one can guess the exact weights, the separation continues to be affected by error in view of the fact that the phase of the signal components is also important but not taken into consideration. The relative phase of the sources components overlapping in the same bin also affects the total energy by interference.

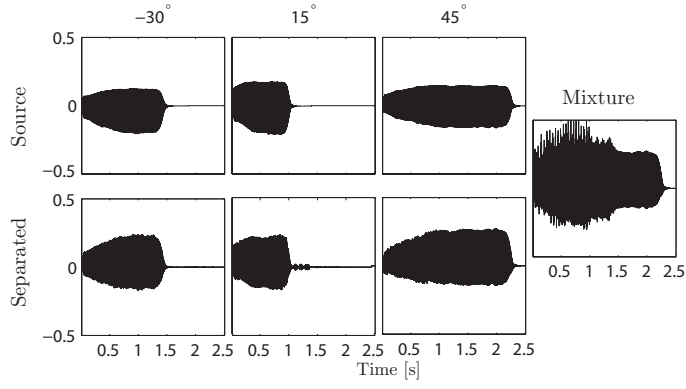


Figure 1.9 Binary mask separation of three trombone sources located at -30°, 15° and 45°: time domain.

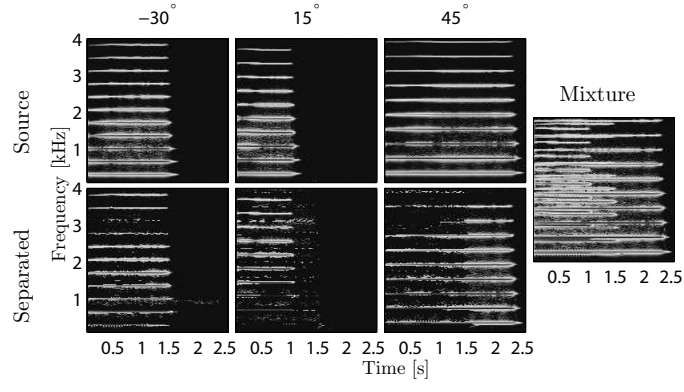


Figure 1.10 Binary mask separation of three trombone sources located at -30°, 15° and 45°: spectrograms (gray scale magnitude).

Gaussian Mixture Model

In theory, in the case of a single source all frequencies should give the same azimuth, exactly corresponding to the source position θ. However, in practice, the violation of the WDO assumption, the presence of noise and estimation errors make things a little more complicated. As a first approximation, we consider that the energy of the source is spread in the power histogram following a Gaussian distribution centered at the theoretical value θ. The Gaussian nature of the distribution is comforted by the well-known Central Limit Theorem as well as practical experiments. In this context, the ideal case is a Gaussian of mean θ and variance 0.

In the case of K sources, we then introduce a model of K Gaussians (K-GMM, order-K

Gaussian mixture model)

$$P_K(\theta|\Gamma) = \sum_{k=1}^{K} \pi_k \ \phi_k(\theta|\mu_k, \sigma_k) \text{ with } \pi_k \geq 0 \text{ and } \sum_{k=1}^{K} \pi_k = 1 \tag{1.29}$$

where Γ is a multiset of K triples $(\pi_k, \mu_k, \sigma_k^2)$ that denotes all the parameters of the model; π_k, μ_k, and σ_k^2 indicate respectively the weight, the mean, and the variance of the k-th Gaussian component described mathematically by

$$\phi_k(\theta|\mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(\theta - \mu_k)^2}{2\sigma_k^2}\right). \tag{1.30}$$

We are interested in estimating the architecture of the K-GMM, that is the number of sources K and the set of parameters Γ, to be able to setup the separation filtering.

Unmixing Algorithm

In the histogram $h(\theta)$, we observe local maxima whose number provides an estimation of the number of sources in the mixture. The abscissa of the k-th local maximum reveals the location θ_k of the k-th source. However, in practice, to avoid spurious peaks, we must deal with a smoothed version of the histogram and consider only significant local maxima – above the noise level. Informal experiments show that the estimated source number and location are rather good. This gives the model order K and a first estimation of the means of the Gaussians (μ_k in Γ). This estimation can be refined and completed – with the variances σ_k^2 and the weights π_k – for example by the EM algorithm.

Expectation Maximization (EM) is a popular approach to estimate parameters in mixture densities given a data set x. The idea is to complete the observed data x with an unobserved variable y to form the complete data (x, y), where y indicates the index of the Gaussian component from which x has been drawn. Here, the role of x is played by the azimuth θ, taking values in the set of all discrete azimuths covered by the histogram. We associate θ with its intensity function $h(\theta)$ (the histogram). The role of y is played by $k \in \{1, \cdots, K\}$, the index of the Gaussian component θ should belong to.

The EM algorithm proceeds iteratively, at each iteration the optimal parameters that increase locally the log-likelihood of the mixture are computed. In other words, we increase the difference in log-likelihood between the current with parameters Γ and the next with parameters Γ'. This log-form difference, noted $Q(\Gamma', \Gamma)$, can be expressed as

$$\begin{aligned} Q(\Gamma', \Gamma) &= \sum_{\theta} h(\theta) \left(\mathcal{L}(\theta|\Gamma') - \mathcal{L}(\theta|\Gamma)\right) \quad \text{with} \\ \mathcal{L}(\theta|\Gamma) &= \log\left(P_K(\theta|\Gamma)\right). \end{aligned} \tag{1.31}$$

We can then reformulate $\mathcal{L}(\theta|\Gamma)$ like this:

$$\begin{aligned} \mathcal{L}(\theta|\Gamma) &= \log\left(\sum_{k} P_K(\theta, k|\Gamma)\right) \quad \text{with} \\ P_K(\theta, k|\Gamma) &= \pi_k \ \phi_k(\theta|\mu_k, \sigma_k). \end{aligned} \tag{1.32}$$

The concavity of the log function allows to lower bound the $Q(\Gamma', \Gamma)$ function using the Jensen's inequality. We can then write

$$Q(\Gamma', \Gamma) \geq \sum_{\theta} \sum_{k} h(\theta) P_K(k|\theta, \Gamma) \log \left(\frac{P_K(\theta, k|\Gamma')}{P_K(\theta, k|\Gamma)} \right) \tag{1.33}$$

where $P_K(k|\theta, \Gamma)$ is the posterior probability, the degree to which we trust that the data was generated by the Gaussian component k given the data; it is estimable with the Bayes rule

$$P_K(k|\theta, \Gamma) = \frac{P_K(\theta, k|\Gamma)}{P_K(\theta|\Gamma)}. \tag{1.34}$$

The new parameters are then estimated by maximizing the lower bound with respect to Γ:

$$\Gamma' = \arg\max_{\gamma} \sum_{\theta} \sum_{k} h(\theta) P_K(k|\theta, \Gamma) \log \left(P_K(\theta, k|\gamma) \right). \tag{1.35}$$

Increasing this lower bound results automatically in an increase of the log-likelihood, and is mathematically easier. Finally, the maximization of Equation (1.35) provides the following update relations (to be applied in sequence, because they modify – update – the current value with side-effects, thus the updated value must be considered in the subsequent relations):

$$\pi_k \quad \leftarrow \quad \frac{\sum_{\theta} h(\theta)\ P_K(k|\theta, \Gamma)}{\sum_{\theta} h(\theta)}, \tag{1.36}$$

$$\mu_k \quad \leftarrow \quad \frac{\sum_{\theta} h(\theta)\ \theta\ P_K(k|\theta, \Gamma)}{\sum_{\theta} h(\theta)\ P_K(k|\theta, \Gamma)}, \tag{1.37}$$

$$\sigma_k^2 \quad \leftarrow \quad \frac{\sum_{\theta} h(\theta)\ (\theta - \mu_k)^2\ P_K(k|\theta, \Gamma)}{\sum_{\theta} h(\theta)\ P_K(k|\theta, \Gamma)}. \tag{1.38}$$

The performance of the EM depends of the initial parameters. The first estimation parameter should help to get around likelihood local maxima trap. Our EM procedure operates as follows:

1. Initialization step
   - initialize K with the order of the first estimation
   - initialize the weights equally, the means according to the first estimation, and the variances with the data variance (for the initial Gaussians to cover the whole set of data):
     $$\pi_k = 1/K, \quad \mu_k = \theta_k, \text{ and } \sigma_k^2 = \text{var}(\theta)$$
   - set a convergence threshold ϵ
2. Expectation step
   - compute $P_K(k|\theta, \Gamma)$ with Equation (1.34)
3. Maximization step
   - compute Γ' from Γ with Equations (1.36), (1.37), and (1.38)
   - if $P_K(\theta|\Gamma') - P_K(\theta|\Gamma) > \epsilon$ then $\Gamma \leftarrow \Gamma'$ and go back to the Expectation step else stop (the EM algorithm has converged).

Finally, to separate the sources, a spatial filtering identifies and clusters bins attached to the same source. Many methods, like DUET, separate the signals by assigning each of the time-frequency bins to one of the sources exclusively. We assume that several sources can share the power of a bin, and we attribute the energy according to a membership ratio – a posterior probability. The histogram learning with EM provides a set of parameters for the Gaussian distribution that characterizes each source. These parameters are then used to parameterize automatically a set of spatial Gaussian filters. In order to recover each source k, we select and regroup the time-frequency bins belonging to the same azimuth θ. We use the parameters issued from the EM-component number k, and the energy of the mixture channels is allocated to the (left and right) source channels according to the posterior probability. More precisely, we define the following mask for each source:

$$M_k(t,f) = P_K(k|\theta(t,f),\Gamma) \tag{1.39}$$

if $10\log_{10}|\phi_k(\theta(t,f)|\mu_k,\sigma_k)| > L_{\mathrm{dB}}$, and 0 otherwise. This mask limits the fact that the tail of a Gaussian distribution stretches out to infinity. Below the threshold L_{dB} (expressed in dB, and set to -20 in our experiments), we assume that a source of interest does not contribute anymore. For each source k, the pair of short-term spectra can be reconstructed according to

$$S_L(t,f) = M_k(t,f)\cdot X_L(t,f), \tag{1.40}$$

$$S_R(t,f) = M_k(t,f)\cdot X_R(t,f). \tag{1.41}$$

Experimental Results

First, we synthesized binaural signals by mixing monophonic source signals filtered through the HRTFs of a given individual corresponding to various azimuths using the HRTF-based technique, then we applied the EM based unmixing algorithm described in the previous section.

A result of demixing is depicted in Figure 1.11 for a two-instrument mixture: xylophone at -55° and horn at -20°; their original spectrograms are shown in Figure 1.12. In the time domain, the xylophone rhythm is respected, its signal looks amplified and its shape is preserved. Perceptively, the demixed xylophone is very similar to the original one. Also, for the horn, we must tolerate some interference effects, and the spectrograms are partly damaged. A portion of energy was absorbed by an unwanted source generated from interferences. We also conducted tests on speech samples. The reconstruction quality was good, much better than for long-distance telephone lines. Figure 1.13 shows the power histogram for the localization of four instruments in a binaural mixture. This histogram, here of size 65, was built using FFTs of $N = 2048$ samples with an overlap of 50%. Figure 1.14 shows the (normalized) Gaussian mixture model associated to this histogram. Of course, the EM algorithm can also be applied on the raw azimuth estimates $\hat{\theta}(t,f)$ instead of the data stored in the histogram.

Multichannel Separation of Overlapping Harmonics

The separation of time-frequency overlapping source components is a very challenging problem. However, in a multichannel context one can use the sensor signals, together with structural cues, e.g., harmonicity, in order to successfully separate the contributions of the various sources to the shared time-frequency components.

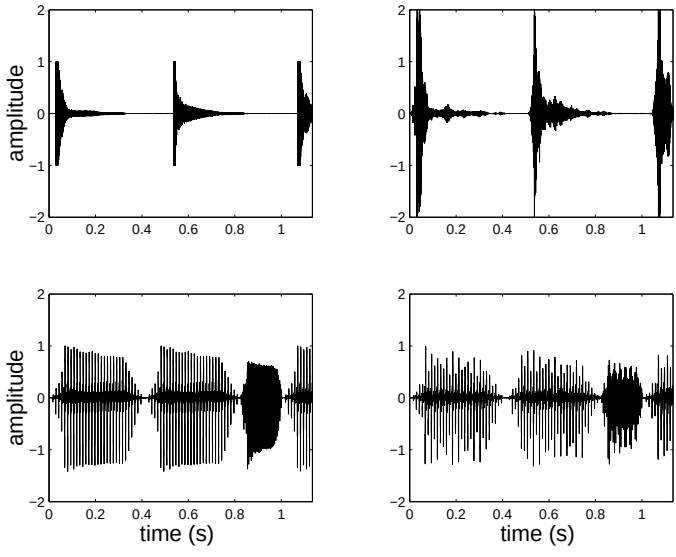


Figure 1.11 Waveforms of the demixtures (on the right, originals being on the left): xylophone (-55°) (top) and horn (30°) (bottom).

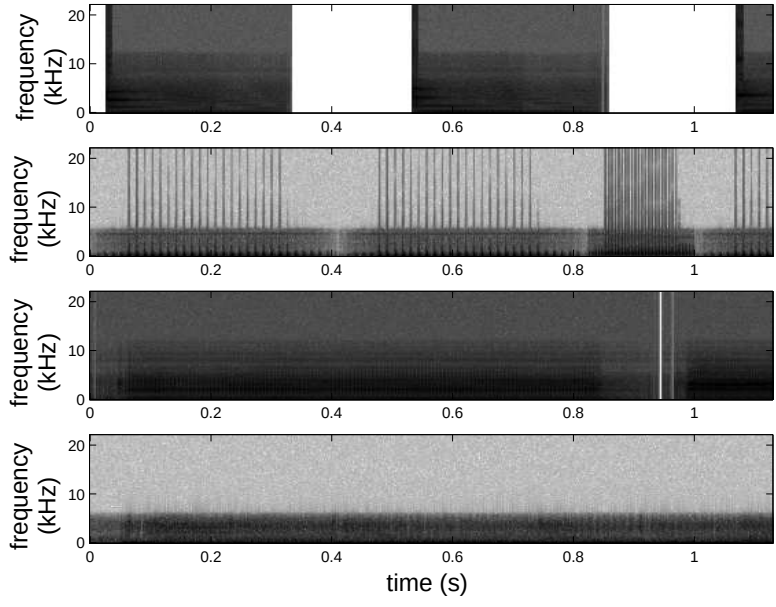


Figure 1.12 Spectrograms of the four sources, from top to bottom: xylophone, horn, kazoo, and electric guitar.

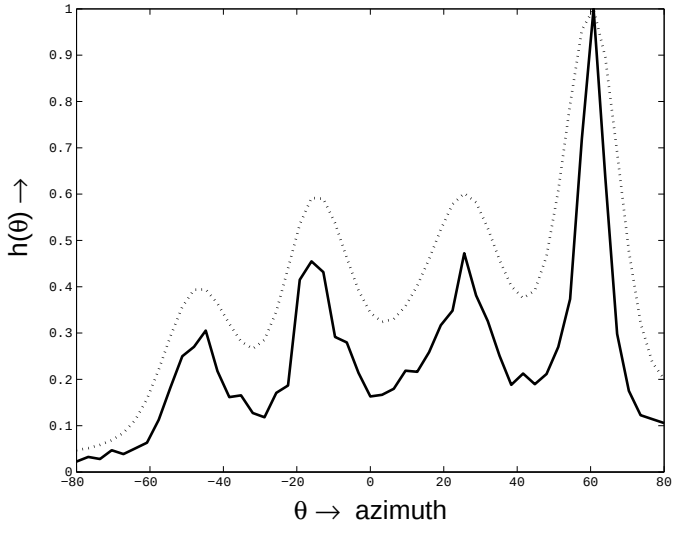


Figure 1.13 Histogram (solid line) and smoother version (dashed line) of the 4-source mix: xylophone at -55°, horn at -20°, kazoo at 30°, and electric guitar at 65°.

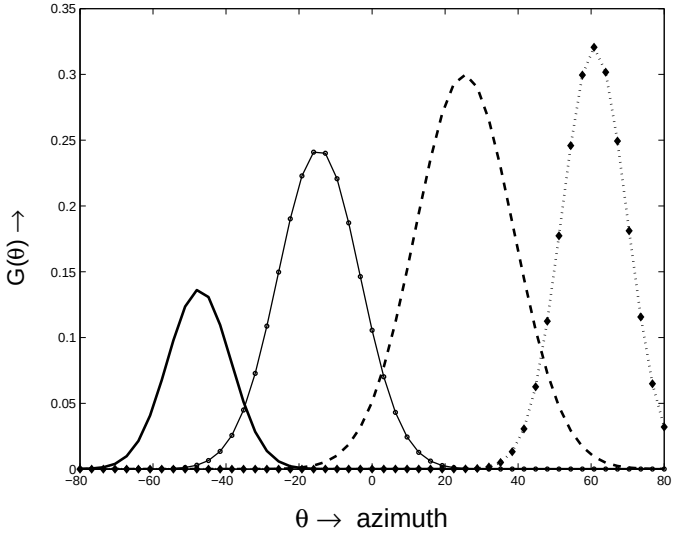


Figure 1.14 GMM (normalized) for the histogram of the 4-source mix.

For each bin of the discrete time-frequency spectrum, let us now consider the observed magnitude in this frequency bin as time goes by. If a single stationary sinusoid with amplitude a and frequency exactly equal to the frequency of one of the analysis bin, the observed magnitude STFT at that bin is proportional to the amplitude of the sinusoid.

When two sinusoids are present at the same frequency, then the resulting signal is also a sinusoid of the same frequency (straightforward when considering the spectral domain). More precisely, the complex instantaneous amplitudes of the sinusoids are adding together as follows:

$$A = a_1 e^{j\phi_1} + a_2 e^{j\phi_2} \tag{1.42}$$

where a_p and ϕ_p are the amplitude and phase of the p-th sinusoid ($p \in \{1, 2\}$). Thus, *via* the Cartesian representation of complex numbers, the corresponding magnitude is

$$a = |A| = \sqrt{(a_1 \cos(\phi_1) + a_2 \cos(\phi_2))^2 + (a_1 \sin(\phi_1) + a_2 \sin(\phi_2))^2}. \tag{1.43}$$

Physically, the addition of two sinusoidal signals, even of same amplitude a_0 ($a_1 = a_2 = a_0$), is ruled by a nonlinear addition law which gives a maximum of $2a_0$ (thus $\approx$ 6dB above the volume of a_0) when the sinusoids are in-phase ($\phi_1 = \phi_2$) and a minimum of 0 when they are opposite-phase ($\phi_1 = \phi_2 + \pi$). One might intuitively think that all the cases in the $[0, 2a_0]$ interval are equiprobable. Not at all! For example, in the case where the initial phases of the sinusoids are independent uniformly distributed random variables over $[-\pi, +\pi)$ we have that $a \approx a_1 + a_2$ is the most probable value for the magnitude, as shown in Figure 1.15. The interference of several sinusoids at the same frequency results in one sinusoid. In this situation, it is quite impossible to separate the two sinusoids without additional knowledge. A possible strategy would be to simply ignore time-frequency regions where this interference phenomenon is likely to occur. However, in the case of musical signals, musicians playing in tempo and in harmony often generate frequency overlap at least for some of the partials. Thus, the separated signals would present large time intervals in which all or some of the partials are muted. When properly exploited, structural cues can help disambiguate overlapping partials and provide heuristics for their separation. In fact, under the assumption of quasi-periodicity of the sources, all the signals are nearly harmonic. Although this may seem a particular case of general inharmonic structure, one can push the methods a little further with simple techniques and minimal side information. In fact, given that at least one of the partials for each source is not completely overlapping in time-frequency with those of the others, one can assume a harmonic structure based on the isolated partials, without additional knowledge of the sources. The overlap of the partials is resolved by assuming that the time envelopes of each partial are similar, with neighboring partial bearing the highest similarity.

A similarity measure was proposed in [HG06], which involved the normalized scalar product of the envelopes:

$$\beta_{p,q} = \frac{\sum\limits_n E_p(n) E_q(n)}{\sqrt{\sum\limits_n |E_p(n)|^2} \sqrt{\sum\limits_n |E_q(n)|^2}} \tag{1.44}$$

where $E_p(n)$ and $E_q(n)$ are the time envelopes of two components detected in time-frequency, each obtained by detecting and tracking in time contiguous zones of non-zero energy in adjacent frequency bins. When the envelopes are quite similar, like the ones of

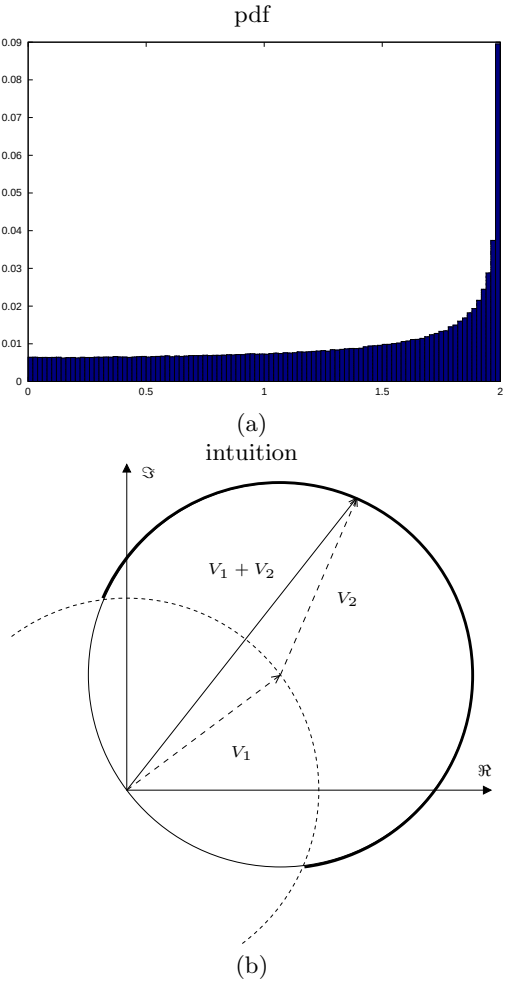


Figure 1.15 Interference of two sinusoids of the same frequency and amplitude 1. The signal resulting from the addition of the these two sinusoids is also a sinusoid with the same frequency, but its amplitude depends on the phases of the initial sinusoids. (a) shows an histogram illustrating the probability density function of this amplitude (for uniform distributions of the phases). The sum of the amplitudes (2) is the most probable value. (b) gives an intuitive justification. It shows that, considering the sum of the two vectors corresponding to the complex amplitudes of the sinusoids in the complex plane (see Equation (1.15)), its norm is more likely to be greater than 1 (bold line).

adjacent harmonics of a single source, the similarity measure $\beta_{p,q}$ is close to one. However, when interference among the sources is present, the similarity measure is much lower than 1.

All the sensor signals can be used in case of multichannel signals. The problem of separating the overlapping partials can be stated as that of estimating a mixing matrix from a known number of N sources contributing to M distinct measurement of each partial. For the problem to have a reliable solution there must be at least as many sensors as there are overlapping partials in that region. In view of the narrow band characteristics of the partials, one can model this mixing process in time-frequency by means of constant complex matrices for each partial [HG06] in each frequency bin. For a two-sensor two-sources case of overlapping partials at frequency bin ω we have:

$$\begin{bmatrix} P_L(n, e^{j\omega}) \\ P_R(n, e^{j\omega}) \end{bmatrix} = \begin{bmatrix} H_{L,1}(e^{j\omega}) & H_{L,2}(e^{j\omega}) \\ H_{R,1}(e^{j\omega}) & H_{R,2}(e^{j\omega}) \end{bmatrix} \begin{bmatrix} S_1(n, e^{j\omega}) \\ S_2(n, e^{j\omega}) \end{bmatrix} \tag{1.45}$$

where $S_i(n, e^{j\omega})$, $i = 1, 2$ represent the STFTs of the original source signals, $H_{L/R,i}(e^{j\omega})$ the frequency responses from source i to left/right sensor channel and $P_{L/R}(n, expj\omega)$ the STFTs of mixed overlapping partials at left and right sensors. The matrix composed by the frequency responses $H_{L/R,i}(expj\omega)$ is the mixing matrix at frequency bin ω.

In order to proceed with separation, first the envelopes of the other non-overlapping partials are computed, which are called the model envelopes. Then, for each candidate mixing matrix, its pseudo-inverse is applied to the sensor mixed partials and the envelopes of the resulting partials are computed. The matrix that gives separated partials with envelopes whose shapes most closely resemble those of the model envelopes is chosen as the estimate of the mixing matrix for that corresponding partial. Thus, the estimate of the mixing matrix is obtained by means of optimization as the one that gives the best match between the separated partials and the model partials according to the criterion (1.44).

Good experimental results are obtained by using the L_1 norm to combine the similarity measures deriving by each partial [Vis04]. An example of separation of overlapping partials from two sources, a violin tone with vibrato and a trombone tone without vibrato, is shown in Fig. 1.16, where one can appreciate how closely the envelopes are recovered together with the fine AM-FM fine structure due to vibrato in the violin tone.

1.3 Source Separation from Single-Channel Signals

Separation of sources from single-channel (monophonic) mixtures is particularly challenging. If we have two or more microphones, we have seen earlier in this chapter that we can use information on relative amplitudes or relative time delays to identify the sources and to help us perform the separation. But with only one microphone, this information is not available. Instead, we must use information about the structure of the source signals to identify and separate the different components.

For example, one popular approach to the single channel source separation problem is to use non-negative matrix factorization (NMF) of the short-term Fourier transform (STFT) power spectogram of the audio signal. This method attempts to identify consistent spectral patterns in the different source signals, allowing these patterns to be associated with the different sources, then allowing separation of the audio signals. As well as NMF, we shall see that methods based on sinusoidal modelling and probabilistic modelling have also been proposed to tackle this difficult problem.

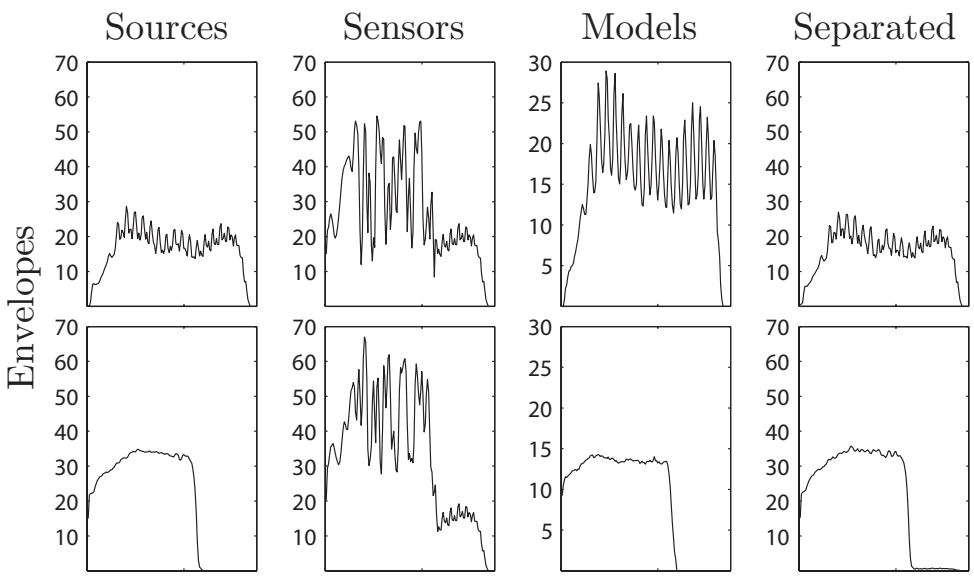


Figure 1.16 Simple example of separation of overlapping partials from two harmonic sources, violin (top row) and trombone (bottom row). Columns from left to right: source envelopes, mixed sources envelopes at left and right sensors, model envelopes extracted from non-overlapping partials at sensor signals, envelopes of separated partials maximizing envelope similarity.

1.3.1 Source separation using Non-negative Matrix Factorization

Suppose that our mixture signal is composed of a weighted mixture of simple source objects each with a fixed power spectrum $\boldsymbol{a}_n$, and where the relative energy of the nth object in the tth frame is given by $s_{nt} \geq 0$. Then the activity of each object has a spectrum of $\boldsymbol{x}_{nt} = \boldsymbol{a}_n s_{nt}$ at frame t. If we then assume that the source signal objects have random phase, so that the spectral energy due to each source object approximately adds in each time-frequency spectral bin, then the spectrum of the mixture will be approximately given by

$$\boldsymbol{x}_t \approx \sum_n \boldsymbol{a}_n s_{nt} \tag{1.46}$$

or in matrix notation

$$\boldsymbol{X} \approx \boldsymbol{A}\boldsymbol{S} \tag{1.47}$$

where $\boldsymbol{X}$ is the mixture spectrogram, $\boldsymbol{A} = [\boldsymbol{a}_n]$ is the matrix of spectra for each source, and $\boldsymbol{S} = [s_{nt}]$ is the matrix of relative source energies in each frame. Since each of the matrices $\boldsymbol{X}$, $\boldsymbol{A}$ and $\boldsymbol{S}$ represent amounts of energy, they are all non-negative.

In single-channel source separation, we only observe the mixture signal with its corresponding non-negative spectrogram $\boldsymbol{X}$. But, since $\boldsymbol{A}$ and $\boldsymbol{S}$ are also non-negative, we can use non-negative matrix factorization (NMF) [LS99] to approximately recover $\boldsymbol{A}$ and $\boldsymbol{S}$ from $\boldsymbol{X}$.

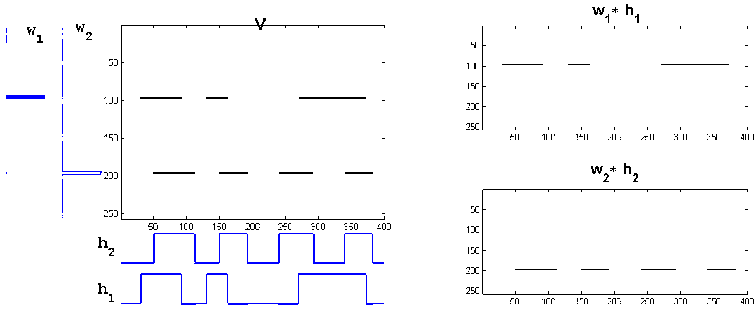


Figure 1.17 Simple example of non-negative matrix factorization.

Non-negative Matrix Factorization (NMF)

In its simplest form, NMF attempts to minimize a cost function $J = D(\boldsymbol{V}; \boldsymbol{WH})$ between a non-negative matrix $\boldsymbol{V}$ and a product of two non-negative matrices $\boldsymbol{WH}$. For example, we could use the Euclidean cost function

$$J_{\mathrm{E}} = D_{\mathrm{E}}(\boldsymbol{V}; \boldsymbol{WH}) = \tfrac{1}{2}\|\boldsymbol{V} - \boldsymbol{WH}\|_{\mathrm{F}}^2 = \tfrac{1}{2}\sum_{pt}(v_{pt} - [\boldsymbol{WH}]_{pt})^2 \tag{1.48}$$

or the (generalized) Kullback-Leibler divergence,

$$J_{\mathrm{KL}} = D_{\mathrm{KL}}(\boldsymbol{V}; \boldsymbol{WH}) = \sum_{pt}\left(v_{pt}\log\frac{v_{pt}}{[\boldsymbol{WH}]_{pt}} - v_{pt} + [\boldsymbol{WH}]_{pt}\right) \tag{1.49}$$

and search for a minimum of J using e.g. a gradient descent method. In fact we can simplify things a little, since there is a scaling ambiguity between $\boldsymbol{W}$ and $\boldsymbol{H}$, so we are free to normalize the columns of $\boldsymbol{W}$ to sum to unity, $\sum_p w_{pn} = 1$. Lee and Seung [LS99] introduced a simple parameter-free "multiplicative update" method to perform this optimization:

$$\begin{aligned} w_{pn} &\leftarrow w_{pn}\sum_t h_{nt}(v_{pt}/[\boldsymbol{WH}]_{pt}) \\ w_{pn} &\leftarrow \frac{w_{pn}}{\sum_p w_{pn}} \\ h_{nt} &\leftarrow h_{nt}\sum_p w_{pn}(v_{pt}/[\boldsymbol{WH}]_{pt}) \end{aligned} \tag{1.50}$$

where this procedure optimizes the KL divergence (1.49).

Fig. 1.17(a) shows a simple example where we have two sources $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ each consisting of a single frequency, with their activations over time given by $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$. Smaragdis and Brown [SB03] applied this type of NMF to polyphonic music transcription, where they were looking to identify the spectra ($\boldsymbol{w}_n$) and activities ($\boldsymbol{h}_n$) of e.g. individual piano notes.

For simple source such as individual notes, we could use the separate products $\boldsymbol{V}_n = \boldsymbol{w}_n\boldsymbol{h}_n^T$ as an estimate of the spectrogram of the original source. We could then transform back to the time domain simply by using the phases in the mixture spectrum as estimates of

the phases of the original sources. However, in practice this approach can be limited, since it assumes that (a) the source spectrum is unchanging over time, and (b) real sources can often change their pitch, which would produce many separate 'sources' from NMF [WP06].

Convolutive NMF

Real sources tend to change their spectrum over time, for example with the high frequency components decaying faster than the low frequency components. To account for this, Virtanen [Vir04] and Smaragdis [Sma04, Sma07] introduced a *Convolutive* NMF approach, known as *non-negative matrix factor deconvolution* (NMFD). Here, each source is no longer represented by a single spectrum, but instead by a typical spectrogram which reflects how the spectrum of the source changes over time.

For NMFD, our model becomes

$$x_{pt} \approx \sum_{n,u} a_{pn}(u) s_{n,t-u} \tag{1.51}$$

which we can write in a matrix form as (Fig. 1.18)

$$\boldsymbol{X} = \sum_{u=0}^{U-1} \boldsymbol{A}(u) \overset{u\rightarrow}{\boldsymbol{S}} \tag{1.52}$$

where the $\overset{u\rightarrow}{\cdot}$ matrix notation means that the matrix is shifted u places to the right

$$[\overset{u\rightarrow}{\boldsymbol{S}}]_{nt} = [\boldsymbol{S}]_{n,t-u}. \tag{1.53}$$

Smaragdis [Sma04] applied NMFD to separation of drum sounds, using it to separate bass drum, snare drum and hi-hat from a mixture.

2D Convolutive NMF

The convolutive approach, NMFD, relies on each drum sound having a characteristic spectrogram that evolves in a very similar way each time that sound plays. However, pitched instruments have a spectrum that is typically composed of spectral lines that change as the pitch of the note changes. To deal with this, Schmidt and Mørup [SM06] extended the (time-) convolutive NMF approach to a 2-dimensional convolutive NMF method, NMF2D. NMF2D uses a spectrogram with a log-frequency scale, so that pitch changes become a vertical shift, while time changes are a horizontal shift (as for NMFD).

The model then becomes

$$x_{pt} \approx \sum_{n,q,u} a_{p-q,n}(u) s_{n,t-u}(q) \tag{1.54}$$

which we can write in a matrix convolution form as (Fig. 1.19)

$$\boldsymbol{X} = \sum_{q=0}^{Q-1} \sum_{u=0}^{U-1} \overset{q\downarrow}{\boldsymbol{A}}(u) \overset{u\rightarrow}{\boldsymbol{S}}(q) \tag{1.55}$$

where the $\overset{q\downarrow}{\cdot}$ matrix notation indicates that the contents of the matrix are shifted q places down

$$[\overset{q\downarrow}{\boldsymbol{A}}]_{pn} = [\boldsymbol{A}]_{p-q,n}. \tag{1.56}$$

Schmidt and Mørup [SM06] used NMF2D to separate trumpet from piano sounds.

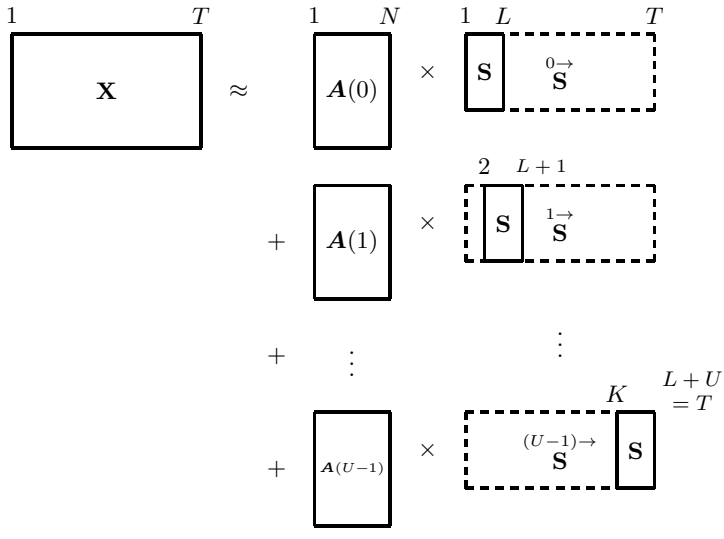


Figure 1.18 Convolutive NMF model for Non-negative Matrix Factor Deconvolution (NMFD)

1.3.2 Structural Cues

Sinusoidal Modeling

Sinusoidal modeling has solid mathematical, physical, and physiological bases. It derives from Helmholtz's research and is rooted in the Fourier's theorem, which states that any periodic function can be modeled as a sum of sinusoids at various amplitudes and harmonically related frequencies. Here we will consider the sinusoidal model under its most general expression, which is a sum of sinusoids (the *partials*) with time-varying amplitudes and frequencies not necessarily harmonically related. The associated representation is in general orthogonal (the partials are independent) and sparse (a few amplitude and frequency parameters can be sufficient to describe a sound consisting of many samples), thus very computationally efficient. Each partial is a pure tone, part of the complex sound (see Figure 1.20). The partials are sound structures very important both from the production (acoustics) and perception (psychoacoustics) points of views. In acoustics, they correspond to the modes of musical instruments, the superpositions of vibration modes being the solutions of the equations for vibrating systems (*e.g.* strings, bars, membranes). In psychoacoustics, the partials correspond to tonal components of high energy, thus very important for masking phenomena. The fact that the auditory system is well adapted to the acoustical environment seems quite natural.

From a perceptual point of view, some partials belong to the same *sound entity* if they are perceived by the human auditory system as a unique sound when played together. There are several criteria that lead to this perceptual fusion. After Bregman [Bre90], we consider

- the common onsets/offsets of the spectral components;

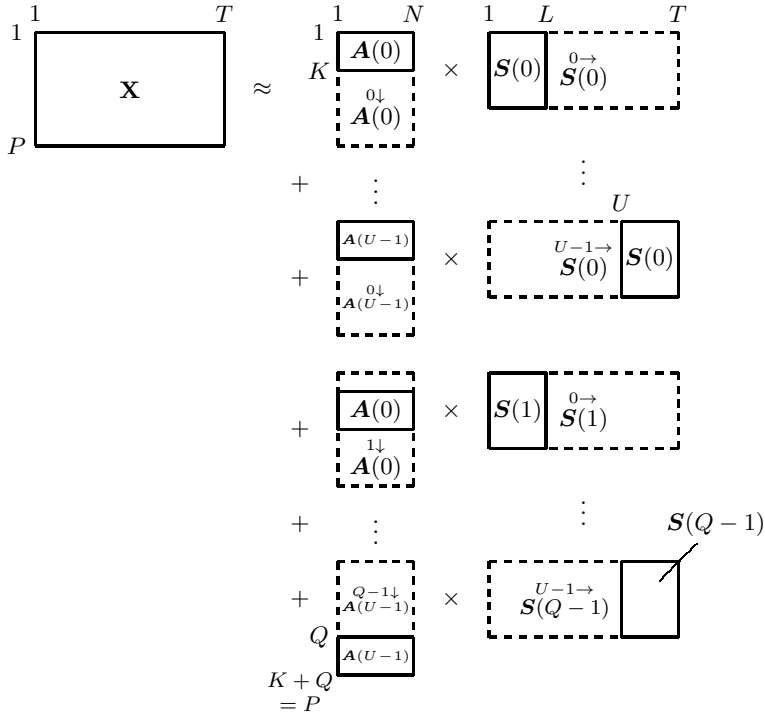


Figure 1.19 Two-dimensional convolutive NMF model (NMF2D)

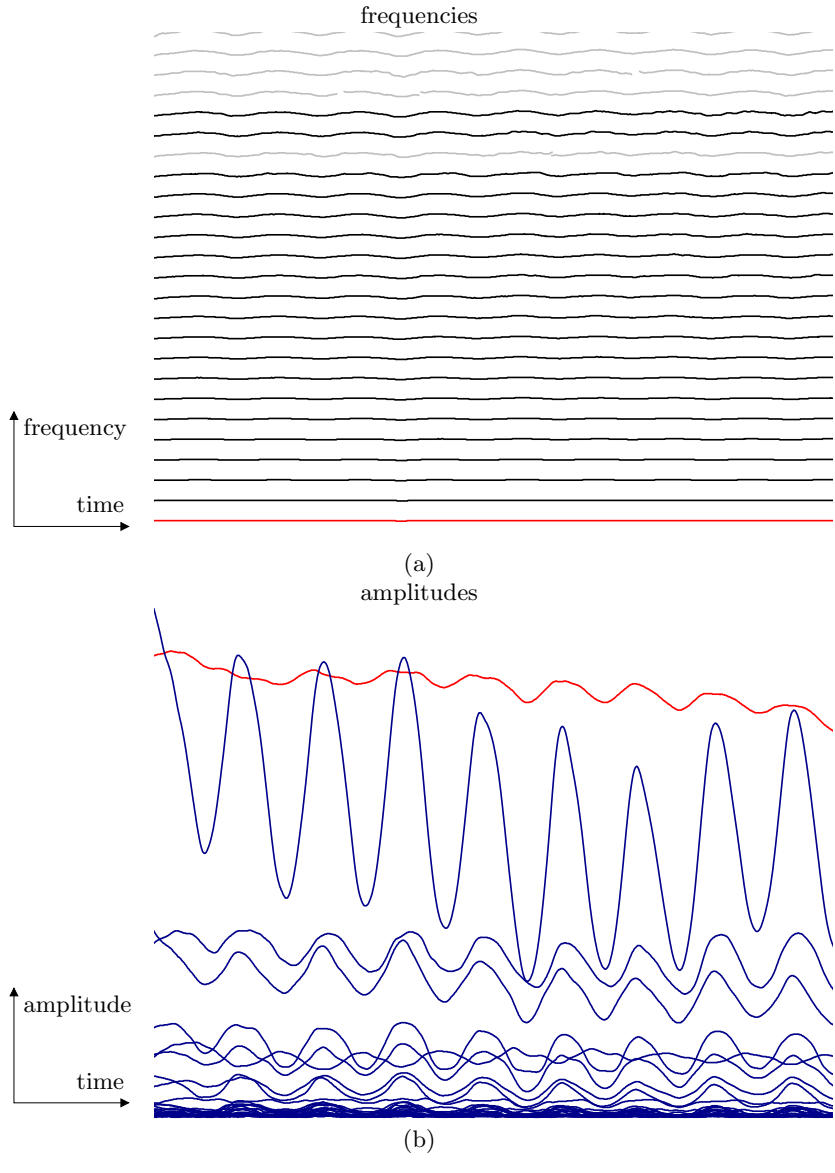


Figure 1.20 The evolutions of the partials of an alto saxophone during $\approx$ 1.5 second. The frequencies (a) and amplitudes (b) are displayed as functions of time (horizontal axis).

- the spectral structure of the sound, taking advantage of harmonic relations;
- the correlated variations of the time evolutions of these spectral parameters;
- and the spatial location, estimated by a localization algorithm.

All these criteria allow us to classify the spectral components. Since this is done according to the perception, the classes we obtain should be the sound entities of the auditory scene. And the organization of these sound entities in time should give the musical structure. Music transcription is then possible by extracting musical parameters from these sound entities. But an advantage over standard music information retrieval (MIR) approaches is that here the sound entities are still available for transformation and resynthesis of a modified version of the music.

The use of each criterion gives a different way to classify the partials. One major problem is to be able to fuse these heterogeneous criteria, to obtain a unique classification method. Another problem is to incorporate this structuring within the analysis chain, to obtain a partial tracking algorithm with multiple criteria that would track classes of partials (entities) instead of individual partials, and thus should be more robust. Finding solutions to these problems are major research directions.

Common Onsets

As noted by Hartmann [Har88], the common onset (birth) of partials plays a preponderant role in our perception of sound entities. From the experiences of Bregman and Pinker [BP78] and Gordon [Gor84], the partials should appear within a short time window of around 30ms (corresponding to a number of γ consecutive frames, see below), else they are likely to be heard separately. Many onset detection methods are based on the variation in time of the amplitude or phase of the signal (see [BDS$^+$05] for a survey). Lagrange [Lag04] proposes an algorithm based on the D measure defined as

$$D[n] = \frac{B[n]}{C[n]} \tag{1.57}$$

$$\text{with} \quad B[n] = \sum_{p=1}^{P} \epsilon_p[n]\bar{a}_p \quad \text{and} \quad C[n] = \frac{1}{2\gamma+1} \sum_{p=1}^{P} \sum_{k=-\gamma}^{+\gamma} a_p[n+k] \tag{1.58}$$

where a_p is the amplitude of the partial p, $\bar{a}_p$ is its mean value, and $\epsilon_p[n]$ is 1 if the partial is born in the $[n-\gamma,n+\gamma]$ interval and 0 otherwise. Thanks to this measure, it seems that we can identify the onsets of notes even if their volume fades in slowly (see Figure 1.21), leading to a better structuring of the sound entities (see Figure 1.22).

Simple and Poly- Harmonic Structures

The earliest attempts at acoustical entity identification and separation consider harmonicity as the sole cue for group formation. Some rely on a prior detection of the fundamental frequency [Gro96] and others consider only the harmonic relation of the frequencies [Kla03]. A classic approach is to perform a correlation of the short-term spectrum with some template, which should be a periodic function, of period F – the fundamental frequency under investigation. Such a template can be built using the expression of the Hann window

$$g_{1,F}(f) = \frac{1}{2}\left(1 + \cos\left(\frac{2\pi f}{F}\right)\right) \tag{1.59}$$

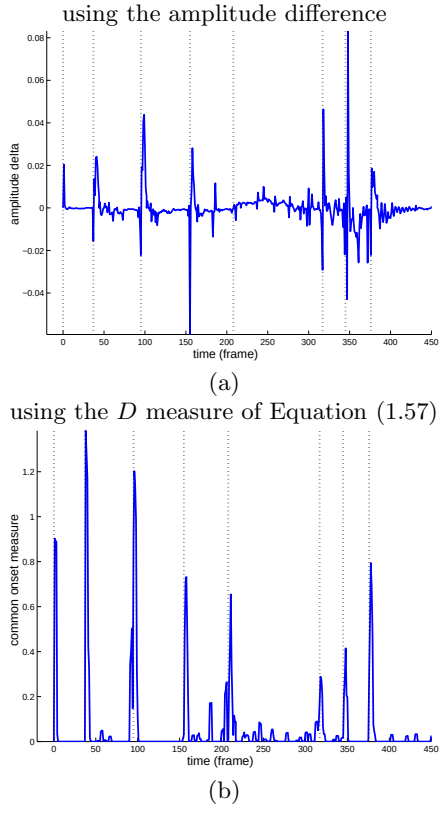


Figure 1.21 Examples of note onset detection using two different measures: (a) the difference in amplitude between consecutive frames and (b) the D measure of Equation (1.57). The true onsets – annotated manually – are displayed as vertical bars. Our method manages to identify correctly the note onsets, even if the volume of the note fades in slowly (see around frame 200).

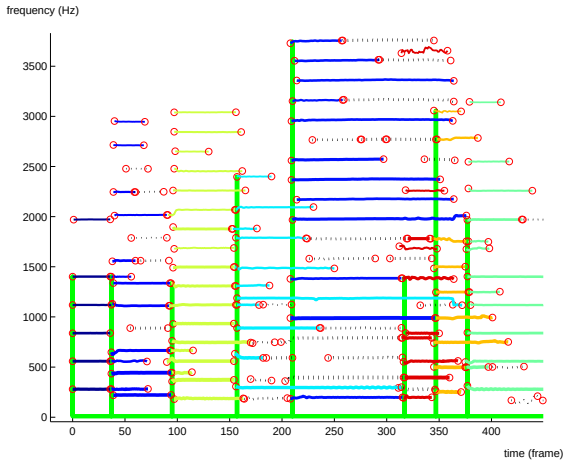


Figure 1.22 Result of the structuring of the partials in sound entities using the common onset criterion, on the example of Figure 1.21.

where f is the frequency and F is the fundamental. However, the problem of these templates is that the width of the template peaks depend on F (see Figure 1.23.a). For a constant peak selectivity, we can propose another function

$$g_{2,F}(f) = g_{1,F}(f)^{-s/\log(g_{1,F}(1))} \tag{1.60}$$

where $s \in (0, 1]$ allows us to tune this selectivity (see Figure 1.23.b). But this template is still too far from the typical structure of musical sounds, whose spectral envelope (in dB) is often a linear function of the frequency (in Hz), see [Kla03]. Thus, we can consider

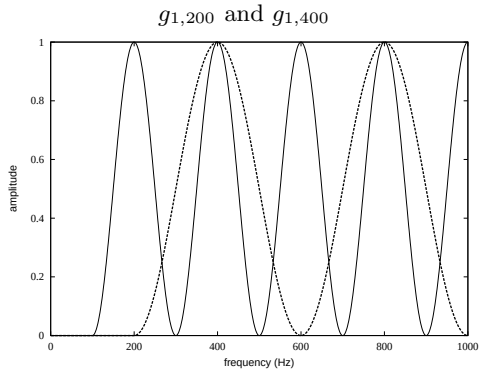
$$g_{3,F}(f) = g_{2,F}(f) \cdot 10^{-d(f-F)} \tag{1.61}$$

where d is the slope of the spectral envelope, in dB per Hz (see Figure 1.23.c). Multi-pitch estimation is still an active research area, and yet our simple approach gives very good results, especially when the s and d parameters can be learned from a database. However, many musical instruments are not perfectly harmonic, and the template should ideally also depend on the inharminicity factor in case of inharmonic sounds.

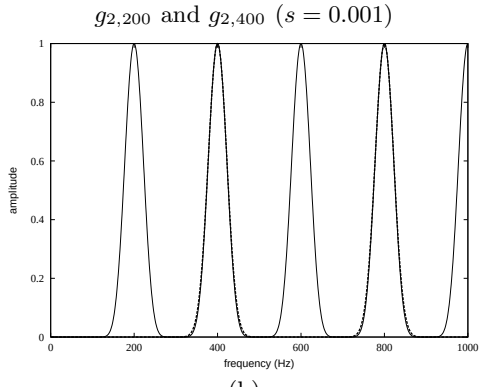
Similar Evolutions

According to the work of McAdams [McA89], a group of partials is perceived as a unique sound entity only if the variations of these partials are correlated, whether the sound is harmonic or not.

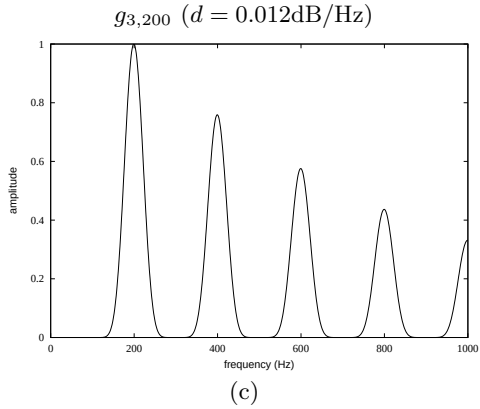
An open problem is the quest for a relevant dissimilarity between two elements (the partials), that is a dissimilarity which is low for elements of the same class (sound entity) and high for elements that do not belong to the same class. It turns out that the auto-regressive (AR) modeling of the parameters of the partials is a good candidate for the design of a robust dissimilarity metric.



(a)



(b)



(c)

Figure 1.23 Three kinds of templates for the extraction of harmonic structures: (a) simple periodic function, (b) modified version for a constant peak selectivity, and (c) modified version for a more realistic spectral envelope.

Let ω_p be the frequency vector of the partial p. According to the AR model, the sample $\omega_p[n]$ can be approximated as a linear combination of past samples

$$\omega_p[n] = \sum_{k=1}^{K} c_p[k]\omega_p[n-k] + e_p[n] \tag{1.62}$$

where $e_p[n]$ is the prediction error. The coefficients $c_p[k]$ model the predictable part of the signal and it can be shown that these coefficients are scale invariant. On contrary, the non-predictable part $e_p[n]$ is not scale invariant. For each frequency vector ω_p, we compute a vector $c_p[k]$ of 4 AR coefficients with the Burg method. Although the direct comparison of the AR coefficients computed from the two vectors ω_p and ω_q is generally not relevant, the spectrum of these coefficients may be compared. The Itakura distortion measure [Ita75], issued from the speech recognition community can be considered:

$$d_{\mathrm{AR}}(\omega_p, \omega_q) = \frac{1}{2\pi} \log \int_{-pi}^{+pi} \left| \frac{C_p(\omega)}{C_q(\omega)} \right| d\omega \tag{1.63}$$

where

$$C_p(\omega) = 1 + \sum_{k=1}^{K} c_p[k] e^{-jk\omega}. \tag{1.64}$$

Another approach may be considered. Indeed, the amount of error done by modeling the vector ω_p by the coefficients computed from vector ω_q may indicate the dissimilarity of these two vectors. Let us introduce a new notation e_p^q, the cross prediction error defined as the residual signal of the filtering of the vector ω_p with c_q

$$e_p^q[n] = \omega_p[n] - \sum_{k=1}^{K} c_q[k]\omega_p[n-k]. \tag{1.65}$$

The principle of the dissimilarity d_σ is to combine the two dissimilarities $|e_p^q|$ and $|e_q^p|$ to obtain a symmetrical one:

$$d_\sigma(\omega_p, \omega_q) = \frac{1}{2}\left(|e_p^q| + |e_q^p|\right). \tag{1.66}$$

Given two vectors ω_p and ω_q to be compared, the coefficients c_p and c_q are computed to minimize the power of the respective prediction errors e_p and e_q. If the two vectors ω_p and ω_q are similar, the power of the cross prediction errors e_p^q and e_q^p will be as weak as those of e_p and e_q. We can consider an other dissimilarity d'_σ defined as the ratio between the sum of the crossed prediction errors and the sum of the direct prediction errors:

$$d'_\sigma(\omega_p, \omega_q) = \frac{|e_p^q| + |e_q^p|}{1 + |e_p| + |e_q|}. \tag{1.67}$$

Lagrange [Lag05] shows that the metrics based on AR modeling perform quite well.

1.3.3 Probabilistic Models

We can also view the source separation problem as a probabilistic inference problem. In outline, using a probabilistic model $p(x|s_1, s_2)$ for the observed mixture x from two sources

s_1 and s_2, and then attempt to find the original sources given the mixture. This is typically done using the maximum a-posteriori (MAP) criterion

$$(\hat{s}_1, \hat{s}_2) = \arg\max_{s_1, s_2} p(s_1, s_2|x) \tag{1.68}$$

using the Bayesian formulation $p(s_1, s_2|x) \propto p(x|s_1, s_2)p(s_1)p(s_2)$ where the sources s_1 and s_2 are assumed to be independent, although other criteria such as the posterior mean (PM) are also possible.

Benaroya et al. [BBG06] use this probabilistic approach, modelling the sources using Gaussian mixture models (GMMs) and Gaussian scaled mixture models (GSMMs). A GSMM is a mixture of Gaussian scaled densities, each of which corresponds to a random variable of the form $g_a = \sqrt{a}g$ where g is a Gaussian with variance σ^2 and a is a non-negative scalar random variable with prior density $p_0(a)$.

They demonstrated their approach on separation of jazz tracks, considering "piano + bass" as one source and drums as another source. The model parameters were trained on a CD containing the separated tracks. This type of CD was originally designed for people to learn how to play jazz, but is also convenient for this type of experiment since it contains separated sources which make a coherent piece of music when mixed together. They found good performance with about 8 or 16 components, although this depended on the particular model used and the estimation method (MAP or PM).

1.4 Applications

In a binaural context, sound source separation has been applied to speech in hearing aids systems for the reduction of the cocktail party effect. The same is true for mobile telephony, where the reduction of environmental noise plays a critical role in the intelligibility of speech. However, even the listening experience of music, both at home and in concert halls, can be enhanced for the listener as interfering environmental noise sources can be eliminated or attenuated.

In a broader single-channel or multi-channel context there are other compelling reasons for desiring to separate sound sources. Traditionally, the listener is considered as a receptor who passively listens to the audio signal stored on various medias (CD, DVD audio, *etc.*) or streamed through the internet. The only modifications that are easy to perform are global to the whole piece, like changing the volume, the tone or adding artificial reverberation. Although new formats such as MPEG Audio Layer 3 (MP3) have changed the way people access to music, the interaction with music is still very limited. However, with the availability of higher computing capabilities, people are more eager to interact with the original media, while the sound is playing. This can be seen for example with the karaoke, where the listener can replace the voice of the original singer. But more freedom and creativity are also possible.

With the techniques presented in this chapter, new ways are available for the identification, separation, and manipulation of the several sound entities (sources) which are perceived by the listener as independent components within the binaural (stereophonic) mix that reaches his/her ears. More precisely, one can find out the sound entities by considering their localization (spatial hearing) and the correlations of their spectral parameters (common onsets, harmonic relations, similar time evolutions). Then, it is possible to apply digital audio effects on each separate sound entity.

This way, listeners are enabled towards an active listening behavior, which entails freedom to interact with the sound in real time during its diffusion. For example, the listener can explore the possibility to change the spatial locations of the individual sound sources, their relative volumes, pitches, and even timbres, as well as their durations, or the rhythm of the music and experiment with these alterations while playing the piece. In other words, by means of sound source separation, the engaging world of mixing and re-editing separate tracks is re-opened to the listner without the need to store or stream the separate tracks, which would require much larger storage or higher data rates.

1.5 Conclusions

In this chapter we have reviewed several methods addressing the problem of source separation in various contexts, such as multichannel, binaural, stereo or mono, vith various degrees of a priori information or assumptions on the sources. To date the results achieved are fairly good considered the complexity and ill-posedness of the problem but near-perfect separation is still beyond reach in most cases. In some applications such as the addition of audio effects or re-spatialization of the sources, near-perfect separation is not a must as artifacts are less udible in the remixed sound. This field has undergone tremendous development in recent years, of which we could only present a partial view, and a considerable amount of work is still under progress.

Acknowledgements

The authors would like to thank Beiming Wang, Harald Viste and Mathieu Lagrange for assistance with some of the figures in this chapter.

Bibliography

[BBG06] L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):191–199, January 2006.

[BDS+05] J. P. Bello, L. Daudet, S.Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, September 2005.

[Bla01] J. Blauert. *Spatial Hearing*. MITpress, 2001.

[BP78] A. S. Bregman and S. Pinker. Auditory Streaming and the Building of Timbre. *Canadian Journal of Psychology*, 32(1):19–31, 1978.

[Bre90] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, MA, 1990.

[BW01] M. S. Brandstein and D. B. Ward, editors. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, New York, NY, 2001.

[FM04] C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Am.*, 116(5):3075–3089, November 2004.

[Gai93] W. Gaik. Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. *Journal of the Acoustical Society of America*, 94(1):98–110, 1993.

[Gor84] J. W. Gordon. *Perception of Attack Transients in Musical Tones*. PhD thesis, Department of Music, Stanford University, California, USA, 1984.

[Gro96] S. Grossberg. *Pitch Based Streaming in Auditory Perception*. MIT Press, 1996.

[GZ10] R. Gribonval and M. Zibulevsky. Sparse component analysis. In P. Comon and C. Jutten, editors, *Handbook of Blind Source Separation*, pages 367–420. Academic Press, Oxford, UK, 2010.

[Har88] W. M. Hartmann. *Auditory Function: Neurobiological Bases of Hearing*, chapter Pitch Perception and the Segregation and Integration of Auditory Entities, pages 623–645. Wiley, New York, USA, 1988. Gerald M. Edelman, W. Einar Gall and W. Maxwell Cowan (Eds.).

[HG06] H. Viste and G. Evangelista. A Method for Separation of Overlapping Partials Based on Similarity of Temporal Envelopes in Multi-Channel Mixtures. *IEEE Trans. on Audio, Speech, and Language Processing*, 14(3):1051–1061, May 2006.

[Ita75] F. Itakura. Minimum Prediction Residual Principle Applied to Speech Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):67–72, 1975.

[JSY98] P. X. Joris, P. H. Smith, and T. C. T. Yin. Coincidence detection in the auditory system: 50 years after Jeffress. *Neuron*, 21:1235–1238, 1998.

[Kla03] A. P. Klapuri. Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness. *IEEE Transactions on Speech and Audio Processing*, 11(6):804–816, November 2003.

[Lag04] M. Lagrange. *Modlisation sinusodale des sons polyphoniques*. PhD thesis, LaBRI, University of Bordeaux 1, Talence, France, December 2004. In French.

[Lag05] M. Lagrange. A New Dissimilarity Metric for the Clustering of Partials Using the Common Variation Cue. In *Proceedings of the International Computer Music Conference (ICMC)*, Barcelona, Spain, September 2005.

[Lin86] W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. i.simulation of lateralization for stationary signals. *Journal of the Acoustical Society of America*, 80(6):1608–1622, 1986.

[LS99] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 21 October 1999.

[McA89] S. McAdams. Segregation of Concurrrents Sounds: Effects of Frequency Modulation Coherence. *Journal of the Acoustical Society of America*, 86(6):2148–2159, 1989.

[Men10] F. Menzer. *Binaural Audio Signal Processing Using Interaural Coherence Matching*. PhD thesis, EPFL Lausanne, Switzerland, 2010.

[NSO09] F. Nesta, P. Svaizer, and M. Omologo. Cumulative state coherence transform for a robust two-channel multiple source localization. In *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation*, pages 290–297, 2009.

[RVE10] M. Raspaud, H. Viste, and G. Evangelista. Binaural source localization by joint estimation of ILD and ITD. *IEEE Trans. Audio, Speech and Lang. Proc.*, 18:68–77, 2010.

[SAM07] H. Sawada, S. Araki, and S. Makino. Frequency-domain blind source separation. In S. Makino, T.-W. Lee, and H. Sawada, editors, *Blind Speech Separation*, pages 47–78. Springer, Dordrecht, The Netherlands, 2007.

[SB03] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, New Paltz, New York, 19-22 October 2003.

[SM06] M. N. Schmidt and M. Mørup. Nonnegative matrix factor 2-D deconvolution for blind single channel source separation. In *Independent Component Analysis and Signal Separation, International Conference on*, volume 3889 of *Lecture Notes in Computer Science (LNCS)*, pages 700–707. Springer, April 2006.

[Sma04] P. Smaragdis. Non-negative matrix factor deconvolution: Extraction of multiple sound sources from monophonic inputs. In *Independent Component Analysis and Blind Signal Separation: Proceedings of the Fifth International Conference (ICA 2004)*, pages 494–499, Granada, Spain, September 22–24 2004.

[Sma07] P. Smaragdis. Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):1–12, Jan. 2007.

[ST97] R. M. Stern and C. Trahiotis. *Binaural and Spatial Hearing in Real and Virtual Environments, Gilkey and Anderson (Eds.)*, chapter 24, Models of Binaural Perception, pages 499–531. Lawrence Erlbaum Associates, 1997.

[VGF06] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469, 2006.

[Vir04] T. Virtanen. Separation of sound sources by convolutive sparse coding. In *Proceedings of the ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA 2004)*, Jeju, Korea, 3 October 2004.

[Vis04] H. Viste. *Binaural Localization and Separation Techniques*. PhD thesis, EPFL, Lausanne, Switzerland, 2004.

[VJA+10] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies. Probabilistic modeling paradigms for audio source separation. In W. Wang, editor, *Machine Audition: Principles, Algorithms and Systems*. IGI Global, Hershey, PA, 2010.

[WB06] D. L. Wang and G. J. Brown, editors. *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley-IEEE Press, Hoboken, NJ, 2006.

[WP06] B. Wang and M. D. Plumbley. Investigating single-channel audio source separation methods based on non-negative matrix factorization. In A. K. Nandi and X. Zhu, editors, *Proceedings of the ICA Research Network International Workshop, 18-19 Sept 2006*, pages 17–20, 2006.

[WS54] R.S. Woodworth and H. Schlosberg. *"Experimental Psychology"*. Holt, 1954.

[YR04] Ö. Yılmaz and S. T. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.